UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE QUÍMICA

GUILHERME ZAINOTTI MIGUEL FAHUR BOTTINO

MÉTODOS ESTATÍSTICOS DE SELEÇÃO DE RESTRIÇÕES "CROSS-LINKING" PARA DETERMINAÇÃO ASSISTIDA DE ESTRUTURAS DE PROTEÍNAS

CAMPINAS

2019

GUILHERME ZAINOTTI MIGUEL FAHUR BOTTINO

MÉTODOS ESTATÍSTICOS DE SELEÇÃO DE RESTRIÇÕES "CROSS-LINKING" PARA DETERMINAÇÃO ASSISTIDA DE ESTRUTURAS DE PROTEÍNAS

Dissertação de Mestrado apresentada ao Instituto de Química da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Química na área de Físico-Química

Orientador: Prof. Dr. Leandro Martínez

O arquivo digital corresponde à versão final da Dissertação defendida pelo aluno Guilherme Zainotti Miguel Fahur Bottino e orientada pelo Prof. Dr. Leandro Martínez.

CAMPINAS

2019

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Química
Erica Cristina de Carvalho Mansur - CRB 8/6734

Bottino, G. Z. M. F., 1992-
B659m Métodos estatísticos de seleção de restrições "cross-linking" para determinação assistida de estruturas de proteínas / Guilherme Zainotti Miguel Fahur Bottino. – Campinas, SP : [s.n.], 2019.

Orientador: Leandro Martínez.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Química.

1. Cross-linking. 2. Proteômica estrutural. 3. Bioinformática. I. Martínez, Leandro, 1979-. II. Universidade Estadual de Campinas. Instituto de Química. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Statistical methods of cross-linking constraints selection for assisted protein structure determination
Palavras-chave em inglês:
Cross-linking
Structural proteomics
Bioinformatics
Área de concentração: Físico-Química
Titulação: Mestre em Química na área de Físico-Química
Banca examinadora:
Leandro Martínez [Orientador]
Leandro Wang Hantao
Marcelo Falsarella Carazzolle
Data de defesa: 19-02-2019
Programa de Pós-Graduação: Química

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0003-1953-1576
- Currículo Lattes do autor: http://lattes.cnpq.br/3423275260374279

BANCA EXAMINADORA

Prof. Dr. Leandro Martínez (Orientador)

Prof. Dr. Leandro Wang Hantao (Instituto de Química - UNICAMP)

Dr. Marcelo Falsarella Carazzolle (Instituto de Biologia - UNICAMP)

A Ata da defesa assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Este exemplar corresponde à redação final da Dissertação de Mestrado defendida pelo aluno Guilherme Zainotti Miguel Fahur Bottino, aprovada pela Comissão Julgadora em 19 de Fevereiro de 2019.

Educar e educar-se, na prática da liberdade,
não é estender algo desde a "sede do saber",
até a "sede da ignorância" para "salvar",
com este saber, os que habitam nesta.

Ao contrário,
educar e educar-se na prática da liberdade,
é tarefa daqueles que sabem que pouco sabem
– por isso sabem algo e podem assim chegar a saber mais –
em diálogo com aqueles que, quase sempre,
pensam que nada sabem, para estes,
transformando seu pensar que nada sabem
em saber que pouco sabem,

possam igualmente saber mais.

(FREIRE, Paulo. Extensão ou comunicação?. Coautoria de Rosisca Darcy de Oliveira. 12. ed. Rio de Janeiro, RJ: Paz e Terra, c2002. 93p. (O mundo hoje, v.24).

Agradecimentos

Antes de mais nada, agradeço a Deus. Embora muitas vezes as religiões tenham regras obscuras, algumas (regras) são claras e cristalinas, e um exemplo é "Deus sobre todas as coisas". A maioria das pessoas desconhece ou ignora isso, mas eu sou um cientista católico. Sou feliz assim, minha religião participa da relação especial que eu tenho com a investigação da natureza sem significar um repositório de ignorâncias ao qual eu recorro quando o ímpeto científico esgota. Sou muito grato a Deus por ter estado sempre acima (vigilante), ao lado (guardião) e à frente (motivador) nos bons e maus momentos, pessoais e profissionais.

Depois, preciso agradecer muito **(muito!)** ao meu orientador, Leandro Martínez. O Leandro é "o orientador que eu quero ser quando crescer". Se orientar é otimizar uma função de muitos parâmetros, desde o bom debate científico, os *insights* ocasionais e a visão inovadora até o treinamento de um pesquisador responsável com prazos e recursos, passando pela formação ideológica e estratégica do acadêmico, eu acredito e confio que o Leandro está muito perto do máximo global.

Sou muito grato também à minha família (Meu pai Sidney, minha mãe Luciana e minha irmã Luiza) por ter - a despeito de todas as vicissitudes - resguardado a estrutura necessária para que a volta pra casa, aos finais de semana, fosse sempre um bom momento. Meus pais fizeram o impossível, penhorando os próprios sonhos pra que eu pudesse conquistar os meus, e suportaram calados e escondidos a esmagadora realidade dos últimos 2 anos para que eu pudesse ter paz no meu mestrado (ainda que eu não tenha tido a mesma cortesia com eles). Não foi em vão. Agradeço também minha avó Magda, meu avô Eduardo, minha avó Iracema e minha tia Mônica, que nunca deixaram a peteca cair, de coração e com muito sacrifício.

Agradeço ao meu glorioso amigo Henrique Caracho, o "último dos moicanos", que me

acompanha desde 2010 na graduação, foi meu veterano, meu agregado, meu colega de quarto e um grande parceiro durante toda a pós-graduação. Ele tem um coração enorme (embora também uma paciência curtíssima) e sempre está "lá", onde quer que "lá" seja.

Agradeço aos indispensáveis André Bina e Bruno Capelas por continuarem me aguentando mesmo depois de uma década de desventuras. Embora as dimensões espaciais e temporais teimem em nos distanciar, saibam que esses encontros na reta final foram essenciais para terminar isso tudo.

Agradeço a todos os titulares e "dinossauros" da saudosa República do Patrão, minha família longe da família, cuja lista de nomes, sozinha, ocuparia mais de meia página. Essa república é meu lar longe do lar, que tanto tempo me acolheu e continua me acolhendo até hoje. Quando as coisas ficaram muito ruins em 2017 e 2018, eles estiveram lá pra mim e eu nunca vou esquecer disso. Agradeço também ao Pedro Ivo por, além de ser o melhor colega de quarto que a pós-graduação me proporcionou, ter segurado o outro lado da corda quando as coisas foram de mal a pior.

Agradeço à Gabriella Silva pelos bons momentos, que me alegraram tanto durante os cinco anos em que estivemos juntos, e também pelos maus momentos, que - numa terapia intensiva - me fizeram mais forte. Obrigado por ter sido a Monica Geller do meu Richard Burke, a Ilsa Lund Laszlo do meu Richard Blane, a Betty Ross do meu Bruce Banner, a Penny Lane do meu William Miller e muitas outras de muitos meus.

Agradeço também às pessoas do nosso grupo, tanto aqueles que já estavam aqui quando eu cheguei (Adriano Ferruzzi, Gabriel Jara, Tayane Honorato, Luciano Censoni, Álvaro Lopez e Antonio Antonelo) quanto aqueles que chegaram depois de mim (Rafael Vicente, Vinicius Piccoli e Diogo Pimenta). Química computacional pode ser uma área um pouco solitária, mas graças a vocês que sempre tiveram paciência pras minhas conversas e eventuais brincadeiras, a solidão foi um pouco menor.

Aprendi também com o Leandro a agradecer à vibrante (à qual finalmente pertenço) comunidade de gente que desenvolve e aperfeiçoa programas e os distribui de graça. Quase tudo na tese foi feito com software *open source* em Linux, desde as modelagens, as análises, a edição do texto em LaTeX, algumas figuras e todos os gráficos. Não fosse por esses softwares, efetivamente nada do que está aqui exceto a análise bibliográfica seria possível, pois a minha bolsa de mestrado é

inferior a dois salários mínimos e não tem taxa de bancada para eu comprar software especializado nenhum. Mesmo assim, eu sou imensamente grato também às Agências de Fomento e centros especializados que me apoiaram da maneira como foi possível.

Agradeço à CAPES pelo apoio ao Programa de Pós-Graduação em Química do Instituto de Química da UNICAMP. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Agradeço ao CNPq e ao Programa de Bolsas Institucionais do PPGQ pelo meu estipêndio de estudos. O trabalho contou com apoio financeiro do CNPq via Bolsa de Mestrado (Processo CNPq 134502/2017-5).

Por fim, esse projeto foi vinculado ao Cepid (Centro de Pesquisa, Inovação e Difusão) CCE&S - *Center for Computing in Engineering and Sciences*, apoiado financeiramente pela FAPESP (Processos 2010/16947-9, 2013/05475-7, 2013/08293-7). O CCE&S é um centro de pesquisa humanamente, intelectualmente e instrumentalmente incrível, do qual espero participar e contribuir por mais muitos anos.

Resumo

O problema da modelagem biomolecular computacional é um dos grandes temas da bioinformática no século XXI. Nesse ínterim, as modelagens *ab initio* baseadas em conhecimento, particularmente, são as metodologias de escolha para o estudo de problemas de fronteira, como proteínas que apresentam pouca homologia nas bases de dados de estruturas conhecidas. Uma das maneiras de se contrapor a algumas mazelas desse tipo de modelagem é auxiliá-la por meio da provisão de dados instrumentais sobre a estrutura proteica em solução, e um dos experimentos que pode ser utilizado para esse fim é denominado Espectrometria de Massas de *cross-linking*, uma técnica que avalia restrições topológicas superficiais de distância. Essa técnica pode gerar uma grande quantidade de dados, dos quais apenas uma pequena porção representa efetivamente as restrições nativas, dando origem ao desafio aqui explorado de seleção e recuperação das restrições adequadas para fornecer como entrada aos algoritmos computacionais.

No presente trabalho, introduzimos um qualificador de restrições baseado num indicador de qualidade clássico da psicometria, denominado coeficiente de correlação ponto-bisserial. Mostramos, para sistemas diferentes, que em quase todos os casos o coeficiente bisserial, adequadamente aplicado, permite a recuperação de restrições mais discriminantes e informativas em relação a um determinado modelo de referência. Recuperações sucessivas, ao longo de um processo iterativo, introduzem vieses incrementais nos conjuntos modelados a ponto de deslocar o espaço conformacional amostrado para regiões mais próximas do que se acredita ser a conformação correta, fornecendo um aumento genérico na qualidade das modelagens.

Mostramos que, dado um número adequado de restrições recuperadas e uma boa ferramenta de seleção de modelos, as restrições recuperadas por meio do BISCORE permitem um aumento significativo de bons modelos gerados. Por meio de um protocolo sugerido que empregue essa metodologia, implementada num pacote de software desenvolvido nesse projeto, foi possível, em alguns casos, aumentar a quantidade de modelos bem-sucedidos por um fator de até 50 vezes, quando se compara à modelagem de partida sem as restrições.

Abstract

The problem of computational biomolecular modeling is one of the major themes of bioinformatics in the 21st century, and knowledge-based *ab initio* modeling, in particular, is the methodology of choice for the study of bleeding-edge problems, such as proteins that present little homology in the databases of known structures. An interesting way to counteract some drawbacks on this type of modeling is to assist it by providing instrumental data on the protein structure in solution, and one of the experiments that can be drawn for this purpose is called Cross-Linking Mass Spectrometry, a technique that evaluates surface topological distance constraints. This technique can generate a large amount of data, of which only a small portion effectively represents the native constraints, giving rise to the challenge hence explored of selecting and recovering the appropriate constraints to provide as input to computational algorithms.

In the present work, we introduce a constraint score based on a classic psychometric quality indicator, called point-biserial correlation coefficient. We show, for different systems, that in almost all cases, properly applied biserial coefficient allows for the retrieval of more discriminating and informative constraints in relation to a given reference model. Successive constraint retrivals, over an iterative process, introduces incremental biases in the modeled sets to the point of shifting the sampled conformational space towards regions closer to what is believed to be the correct conformation, providing a general increase in modeling quality.

We show that, given an adequate number of recovered constraints and a good model selection tool, the constraints retrieved through BISCORE allow for significant increase in the amount of satisfactory models. Through a protocol that employs this methodology, implemented in a software package developed in this project, it was possible in some cases to increase the number of successful models by a 50-fold, when compared to preliminary unconstrained modeling.

Lista de Abreviaturas

SAXS	Espalhamento de raios-X a pequenos ângulos
DRX	Difração de raios-X
CryoEM	Criomicroscopia eletrônica
RMN	Ressonância Magnética Nuclear
XL-MS	Em inglês, *Cross-Linking Mass Spectrometry*, ou Espectrometria de Massas de *Cross-Linking*.
PolII-TFIIF	Complexo RNA Polimerase II - Fator de Transcrição IIF
XLs	*Cross-links*
CASP	Em inglês, *Critical Assessment of Structure Prediction*, ou Avaliação Exigente de Predição Estrutural, uma competição internacional de modelagem biomolecular
REF	Em inglês, *Rosetta Energy Function*, Função Energia do Rosetta
SALBIII	Proteína Epóxido hidrolase/ciclase de síntese da Salicinomicina
HSA	Proteína Albumina do Soro Humano
MMC	Metropolis-Monte Carlo
CB	Carbono β, referente a cada aminoácido na sequência primária
LOVO	Em inglês, *Low Order-Value Optimization*, ou otimização do menor valor ordenado
PDB	*Protein Data Bank*
RMSD	Em inglês, *Root-mean-square deviation*, ou Raiz do Desvio Quadrático Médio

GDT Em inglês, *Global Distance Test*, ou Teste Global de Distância

FREQ Frequentista

BISCORE-CRYS Coeficiente bisserial empregando como variável contínua a similaridade (TM-score) em relação à estrutura cristalográfica

BISCORE-BEST Coeficiente bisserial empregando como variável contínua a similaridade (TM-score) em relação ao melhor modelo gerado

BISCORE-PROQ3D Coeficiente bisserial empregando como variável contínua a similaridade (TM-score) em relação ao modelo de maior ProQ3D-tmscore

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Modelar é preciso!

A palavra *proteína* gera, na maioria dos cientistas naturais, sensações mistas de familiaridade e estranhamento. Conforme é de conhecimento geral, estão presentes na condução de praticamente todos os processos biológicos, não só no interior das células dos organismos como também na própria composição dessas células [1]. No entanto, o estranhamento deriva de um comportamento extremamente curioso dessas moléculas: elas só são capazes de exercer sua função se estiverem em uma conformação específica – denominada conformação nativa –, que é única para cada sequência de aminoácidos [2].

Esse fenômeno forja uma relação bilateral entre forma e função das proteínas, de maneira que a elucidação de modelos estruturais de proteínas figura entre os problemas multidisciplinares de maior relevância global, abarcando esforços combinados de diferentes áreas do conhecimento num esforço constante de aprimorar a eficácia e eficiência das diferentes técnicas utilizadas na aproximação desse problema. Sua importância é potencializada à medida que se constata que, ao longo das últimas décadas, à proporção que as técnicas de sequenciamento genômico e proteômico foram se desenvolvendo, ampliou-se a capacidade dos cientistas em geral de descobrir e sequenciar novas proteínas, gerando um hiato significativo entre a descoberta e a efetiva caracterização (ou seja, a descrição funcional, e, portanto, estrutural do estado nativo) dessas moléculas [3].

1.1 O Problema da Modelagem Molecular

Tomando como ponto de partida uma proteína qualquer para a qual é conhecida a sequência primária, pode-se enunciar resumidamente que o problema de modelagem computacional

consiste em, por meio de um conjunto de informações - de diferentes naturezas, seja instrumentais, teóricas ou orientadas por dados - e pressupostos - estes, predominantemente físico-químicos-, através de softwares e algoritmos, propor a estrutura tridimensional (ou uma pequena coleção de estruturas tridimensionais) que reflita a estrutura real da proteína [4].

É importante desde já entender que a modelagem biomolecular computacional (por vezes também denominada *"in silico"* ("em silício"), fazendo referência ao principal elemento componente dos microprocessadores, em paralelo aos termos *"in vitro"* e *"in vivo"* comuns nas ciências naturais) não necessariamente substitui as técnicas de aferição instrumental da estrutura proteica. Na realidade, atua mais de forma a auxiliar ou complementar essas técnicas, advindo como uma possível solução às suas limitações. Desse ponto de vista, cabe à computação e à ciência de dados a árdua tarefa de preencher as lacunas e propor uma estrutura onde a qualidade dos dados experimentais em si já é impeditiva de fazê-lo por conta própria.

Por exemplo, no caso da cristalografia por DRX, existe o problema da dificuldade de se cristalizar uma proteína, principalmente aquelas que estão localizadas em ambientes químicos complexos de se reproduzir (como proteínas transmembrana ou intranucleares). Muitas delas requerem o uso de agentes de nucleação específicos e geralmente geram difratogramas que permitem conclusões apenas parciais sobre a estrutura ou, ainda, em que a estrutura cristalizada foi fixada em uma conformação biologicamente e funcionalmente irrelevante [5]. As técnicas de SAXS e Cryo-EM têm como premissa se contrapôr a esses problemas da cristalografia convencional, permitindo a análise de proteínas em solução, mas com o custo de uma redução drástica de resolução. Portanto, combinando limitações teóricas e instrumentais, essas técnicas não produzem ainda dados capazes de, por conta própria, propôr modelos completamente atomísticos de proteínas [6]. No caso da investigação por RMN, o problema é mais de ordem técnica: uma vez que a resolução diminui com o aumento do tamanho da proteína, sistemas com mais de 100kDa requerem equipamentos de altíssima frequênca (acima de 600MHz) e protocolos multidimensionais com longos tempos experimentais para gerar dados minimamente utilizáveis [7].

Ao mesmo tempo em que é importante reduzir o esforço no sentido de estreitar a lacuna entre sequência e estrutura, percebe-se que a despeito de muitos progressos ao longo de mais de 40 anos dessa linha de pesquisa, ainda não foi fundamentada uma metodologia que seja universal e transferível, ou seja, de eficácia reprodutível para diferentes sistemas. Isso porque,

embora o enunciado do problema seja simples, ele está permeado de decisões estratégicas que tanto restringem quanto refletem o tipo de sistema proteico, a qualidade das informações a serem utilizadas, o investimento computacional envolvido e principalmente o tipo de resultado que se deseja obter. Algumas dessas decisões são, por exemplo:

- Qual o nível de escala que será empregado no modelo tridimensional?
- Como será criada ou sugerida a conformação inicial (ou de partida) para a modelagem?
- A presença do solvente é importante? Se for, como ele será descrito?
- Existe uma estrutura cristalográfica, ainda que de baixa resolução e com lacunas? Há homólogos conhecidos?
- Quais são os valores numéricos que definem a progressão de uma modelagem? Qual é o critério de parada de uma modelagem?

Entretanto, se for possível selecionar a pergunta mais decisiva que se deve responder, provavelmente ela pode ser sintetizada em: **qual é a estratégia geral de modelagem?** A resposta dessa pergunta, uma vez estabelecida a hipossuficiência dos dados apenas instrumentais para a determinação estrutural, apresenta-se no tipo de sistema, e na possibilidade de aplicação de todas as informações coletadas sobre ele.

1.2 Estratégias de Modelagem e o Paradoxo de Levinthal

Quando as primeiras estratégias computacionais de investigação de proteínas surgiram, na década de 1970, alguns importantes trabalhos já haviam sido realizados para tentar compreender o comportamento da naturação e denaturação proteica em diferentes meios. Atribui-se a Cyrrus Levinthal, entre os anos de 1968-69, uma série de experimentos nesse ínterim que levou ao estabelecimento do que posteriormente ficou conhecido como "Paradoxo de Levinthal", que pode ser enunciado da seguinte forma:

> Em contraste com o fato de que a maioria das proteínas, por conta do seu tamanho, tem acesso a um espaço conformacional exorbitantemente grande (por exemplo, uma proteína de 150 aminoácidos, onde cada um possui três conformações possíveis, apresenta 3^{150} graus de liberdade!), a transição entre os estados denaturado e

> enovelado de uma proteína ocorre naturalmente em escalas de tempo surpreendentemente pequenas, da ordem de minutos [8, 9].

À época, o próprio Levinthal propôs que isso era uma consequência da maneira como a própria proteína, por diferentes razões, amostrava seletivamente apenas pequenas porções desse espaço conformacional [9].

Muitas hipóteses foram traçadas contemporaneamente para tentar explicar esse fenômeno. Uma dessas hipóteses, creditada a Christian Anfinsen e conhecida como Hipótese Termodinâmica [10], dizia que a conformação - ou conjunto de conformações - que existe em solução e nas condições funcionais da proteína (que pode ser chamado de estado nativo), na verdade, equivale a um mínimo de energia livre num "funil configuracional" [11]. A projeção da superfície do "funil" representaria um conjunto de conformações e a profundidade do "funil" representaria a energia livre da configuração. Essa figura do "funil" de energia livre é até hoje empregada em obras didáticas de bioquímica a nível de graduação [1, 2, 12].

1.2.1 Esforços iniciais

À proporção que a hipótese termodinâmica ganhava força nos anos 70 e 80, métodos físico-químicos de cálculo de Energia Livre foram desenvolvidos com embasamento físico, e campos de força clássicos foram criados e otimizados para o estudo de proteínas em simulações de Dinâmica Molecular. A linha de pesquisa é considerada oficialmente inaugurada quando, em, 1975, Levitt simula um movimento de renaturação parcial do inibidor da tripsina pancreática bovina [13]. Em menos de quinze anos, seriam lançados três campos de força que figuram, até hoje, entre os mais populares do mundo, a saber: o campo GROMOS *("GROningen MOlecular Simulation")*, lançado em 1978 e documentado oficialmente em sua segunda versão, GROMOS87, em 1987 [14]; o campo CHARMM *("Chemistry at HARvard Macromolecular Mechanics")*, lançado oficialmente em 1983 [15]; e o campo OPLS, *("Optimized Potentials for Liquid Simulations")* em 1988 [16].

Embora os paradigmas e premissas da Dinâmica Molecular não pertençam ao escopo desse projeto, é importante fornecer uma pequena introdução sobre os mesmos. Esse tipo de simulação se baseia em observar a evolução dinâmica de um certo sistema de partículas ao longo de um determinado tempo, e, para alguns tipos de sistemas, utilizar essa evolução para calcular

propriedades macroscópicas. Para computar as trajetórias dos átomos nos sistemas, a cada passo de tempo, o potencial é calculado em diferentes pontos do sistema por meio do campo de forças escolhido e dentro das condições de contorno da simulação. Os campos de forças clássicos são extensas funções potencial, que recebem distâncias, ângulos e diedros entre átomos e retornam valores de Energia. Posteriormente, as partículas são perturbadas por meio da solução das equações de Newton, onde estima-se numericamente o gradiente desse potencial, as novas posições são computadas e o ciclo se reinicia para o próximo passo de tempo.

No caso de proteínas, na prática, experimentos de modelagem envolveriam partir de diversas configurações iniciais aleatoriamente desenoveladas e tentar atingir uma configuração enovelada dentro de um tempo de simulação razoável. Os campos de força mais primitivos, como aqueles que estavam em voga nos anos 70, já obtinham resultados excelentes no refino de estruturas, mas, conforme notado pelo próprio Levinthal [9], essas técnicas, à época, eram meramente capazes de trazer uma estrutura proteica ao mínimo local mais próximo, sendo incapazes de alcançar um estado enovelado a partir de uma estrutura estendida. As mazelas desse processo, além dos percalços matemáticos, se resumem a quatro principais pontos [17]:

- a escala de tempo para o enovelamento de proteínas, da ordem de alguns minutos, é muito maior que a escala de tempo de simulações de MD atomísticas (que utilizam passos de tempo da ordem de 1 fs e se estendem no máximo a alguns microssegundos), sendo essa limitação de ordem computacional e tornando incrementos de escala impossíveis;
- a maioria dos campos de força, mesmo aqueles mais modernos, não tem a precisão necessária para lidar com as nuances do enovelamento proteico;
- muitas vezes o enovelamento envolve a formação e rompimento de ligações químicas, fenômenos impossíveis de descrever com simulações de DM puramente clássicas; embora já existam métodos multiescala tais quais QM/MM *(Quantum Mechanics/Molecular Mechanics)*, que permitem a simulação de fenômenos eletrônicos, a introdução de uma região quântica na simulação agrava ainda mais o problema da escala de tempo;
- conforme a proteômica estrutural foi evoluindo, descobriu-se que apenas uma pequena parcela das proteínas se enovela por completo de forma espontânea em solução;

Por isso, para a modelagem *de novo* de proteínas, as estratégias de DM nunca

chegaram a ser efetivamente cogitadas como uma solução singular ou definitiva, uma situação que apenas na última década começou a ser modificada com o avanço tecnológico da capacidade computacional e estratégias de paralelização massiva [18]. Isso abriu margem ao desenvolvimento de outras estratégias, que ao invés de tentar acompanhar a dinâmica de enovelamento, utilizam estratégias mais agressivas de amostragem combinadas a características de estruturas proteicas já conhecidas, com campos de força baseados em informações de outras proteínas. A esse tipo de modelagem, dá-se o nome de modelagem baseada em conhecimento.

1.2.2 Modelagem baseada em conhecimento

Existem duas categorias principais de modelagem baseada em conhecimento, a modelagem por homologia e a modelagem *ab initio*. No primeiro caso, a modelagem é baseada principalmente em zonas conservadas de proteínas homólogas, às quais uma sequência de aminoácidos é encaixada e posteriormente refinada. Didaticamente, é possível pensar que, na modelagem por homologia, a sequência primária é costurada (em inglês, *"threaded"*) num conjunto de pontos que vêm de proteínas similares, como se a primeira fosse uma linha e o segundo uma trama de tecido. Uma vez que as configurações amostradas, as perturbações sugeridas e energias calculadas são muito dependentes dos homólogos, a qualidade geral da modelagem também o é. Portanto, essa técnica só fornecerá bons resultados quando houver muitos homólogos de boa qualidade - com grandes zonas bem conservadas - nas bases de dados proteicas. Já na modelagem *ab initio*, o uso da informação é mais sutil e indireto. Quando é impossível contar com grandes zonas conservadas, utilizam-se fragmentos polipeptídicos ou oligopeptídicos, métodos de amostragem mais combinatórios e parâmetros dos campos de forças mais baseados em estatística descritiva das bases de conhecimento.

Independentemente da estratégia, é importante entender que o produto final da modelagem é uma estrutura, ou um pequeno conjunto de estruturas, que procura representar o estado nativo ou quase nativo. Numa modelagem baseada em conhecimento, e este projeto não é exceção, geralmente são gerados alguns milhares de modelos, dentre os quais os candidatos à solução são selecionados. Justamente por isso existe um interesse genuíno em ampliar a qualidade da modelagem para que forneça uma proporção maior de modelos com alta confiabilidade (medida que será explorada mais à frente): primeiro porque aumenta a probabilidade de que um

subconjunto dos modelos gerados reflita o estado real; segundo, porque permite diminuir o dispêndio computacional, gerando menos modelos e permitindo, por exemplo, explorar o potencial iterativo das modelagens.

1.3 Modelagem Computacional Assistida

Efetivamente, os pressupostos já mencionados de Levinthal e Anfinsen foram a força motriz de diversos trabalhos motivados por procurar conectar o espaço configuracional das proteínas a uma interpretação cinética e termodinâmica do processo de enovelamento. Entretanto, outro aspecto importantíssimo é que esses pressupostos são responsáveis pela principal menção àquele que ainda é o grande obstáculo, ou gargalo, da predição estrutural *de novo* de proteínas, que é a amostragem conformacional [19].

1.3.1 Campos de força e sua contribuição

Para contornar dificuldades de amostragem, os protocolos de modelagem de proteínas foram se desenvolvendo para que a contribuição das energias estimadas - calculadas por meio dos campos de forças clássicos - fossem se tornando cada vez mais precisas, abrangentes e personalizáveis, pois delas dependeria limitar o espaço conformacional e tornar a modelagem de proteínas mais factível. De fato, se os campos de força efetivamente introduzirem um viés adequado - pendente à conformação real - à busca conformacional, os tempos necessários para alcançar o enovelamento se tornam mais práticos.

Enquanto isso é verdade, uma série de estudos já mostrou que, quando se trata de proteínas grandes (com mais de 100 aminoácidos em um só domínio), os pequenos erros dos campos de força - que geralmente são parametrizados por cálculos teóricos - são propagados [20] até um ponto em que os erros sistemáticos e aleatórios são impeditivos de alcançar uma modelagem adequada [21]. À proporção que alguns pesquisadores se dedicam a ampliar a precisão dos campos de força (com novas parametrizações ou otimizações *post hoc*), muitas vezes, mesmo com refinamentos, alcançar o estado nativo com um campo de forças genérico permanece um desafio. Felizmente, existe outra maneira de se contrapor a esse obstáculo amostral: uma parcela

da comunidade científica se preocupa em personalizar esses campos de força genéricos, introduzindo novos termos que fogem à descrição lugar-comum dos campos de forças clássicos.

Essa personalização pode ocorrer, por exemplo, utilizando dados sistema-específicos, experimentais, de investigação instrumental da proteína em solução. A esse tipo de abordagem do problema dá-se o nome de modelagem assistida ou integrativa [22, 23]. A ideia é que agregando aos termos comuns dos campos de força clássicos outros termos, derivados de aferições experimentais, e atribuindo pesos competitivos a esses novos termos, é possível introduzir mais informação enviesante do espaço conformacional na modelagem. Esses dados costumam ser obtidos a partir das técnicas instrumentais mais comuns na proteômica, como espectroscopias e microscopias.

1.4 Espectrometria de Massas de *Cross-Linking*

Uma dessas técnicas, ainda relativamente jovem, é a chamada Espectrometria de Massas de Cross-Linking (XL-MS), a mais proeminente técnica de proteômica estrutural por Espectrometia de Massas [24]. Nela, são empregados reagentes denominados *cross-linkers*, que são capazes de se conjugar quimicamente a uma proteína por meio de ligações covalentes. A maioria dessas moléculas possui em sua estrutura grupos funcionais de reatividade seletiva, que caracterizam sua utilidade, espaçados por uma cadeia alifática comprida, que caracteriza seu comprimento. Existem também casos onde os cross-linkers apenas promovem a conexão covalente (formação de aduto) entre resíduos, sem introduzir uma cadeia espaçadora.

A ideia por trás dos *cross-links* entre proteínas é razoavelmente antiga na comunidade científica de bioquímica estrutural. Essa técnica foi, durante anos, utilizada para identificar complexos proteicos em experimentos de eletroforese em gel. No primeiro trabalho em que se dá conta de seu emprego de maneira similar ao protocolo moderno, publicado em 1974, *cross-links* permitiram identificar as proteínas vizinhas aos ribossomos de E. coli [25].

Quase três décadas decorreram até o primeiro trabalho que apresenta o acoplamento dessa técnica com a Espectrometria de Massas, contexto em que foram utilizadas para determinar a topologia de um poro nuclear em levedura [26]. Esse trabalho é um grande marco no uso de XL-MS para proteômica estrutural, pois a partir desse estudo, anos depois, foi possível propôr uma estrutura para outra porção do mesmo poro nuclear [27], do qual só existia uma microscopia eletrônica de

baixa resolução com quase uma década de idade [28]. A partir daí, ambas as técnicas começam a ser desenvolvidas e empregadas em conjunto [29], bases de dados e metodologias de análise são criadas, e trabalhos são publicados empregando XL-MS na interpretação e complementação de dados de outras técnicas instrumentais [30, 31, 32].

Em 2010, publica-se finalmente o que é considerado o primeiro artigo de modelagem computacional assistida por XL-MS, um refinamento da estrutura para o complexo proteico PolII-TFIIF [33]. No final do resumo desse trabalho, os autores escrevem *"This work establishes cross-linking/MS as an integrated structure analysis tool for large multi-protein complexes"*.

1.4.1 Tipos de *cross-links* empregados nesse projeto

Existem diversas categorias de *cross-links* que podem ser investigados por meio da técnica de XL-MS. Nesse trabalho, toda a parte experimental foi realizada por nossos colaboradores no grupo de pesquisa do Prof. Dr. Fabio Gozzo (Dalton MS Lab, Instituto de química, Unicamp), e detalhes do protocolo utilizado e das reações orgânicas envolvidas podem ser obtidos em [34]. No entanto, cabe aqui um breve sumário dos tipos de conexão observados.

1. Conexão entre dois resíduos básicos *(DSS cross-link)*

   Pares observáveis: Lys-Lys, Lys-Ser, Ser-Ser, Met(N terminal)-Lys

   Nesse protocolo, é utilizada uma molécula denominada DSS (suberato de succinimidila) ou, no caso do seu produto de sulfonação, mais solúvel em água, BS3 (suberato de bissulfosuccinimidila). Sua estrutura pode ser observada na Figura 1.1. Essa molécula possui duas carbonilas ativadas para substituição nucleofílica, devido ao fato da N-hidroxi-succinimida ser um excelente grupo abandonador. Portanto, essas carbonilas sofrem reações de amidação por meio do ataque nucleofílico das aminas nas cadeias laterais de Lisinas e Serinas;

2. Conexão entre dois resísuos ácidos *(Diamine cross-link)*

   Pares observáveis: Glu-Glu, Asp-Asp, Glu-Asp

   Nesse protocolo, os carboxilatos nas cadeias laterais de Ácidos Glutâmicos e Ácidos Aspárticos são inicialmente ativados para amidação por meio do emprego de uma carbodiimida seguida

de hidroxibenzenotriazol. Ao final desse tratamento, os carboxilatos são convertidos a ésteres de benzenotriazol, que é um excelente grupo abandonador e, portanto, torna as carbonilas dos Ácidos Glutâmicos e Ácidos Aspárticos ativadas para substituição nucleofílica. A adição de uma diamina de cadeia média pode, então, conectar essas carbonilas de resíduos ativados.

3. Conexão entre um resíduo ácido e um resíduo básico *(Zero-Length cross-link)*

   Pares observáveis: Asp-Lys, Asp-Ser, Glu-Lys, Glu-Ser

   Essa conexão, que é um subprocesso da conexão do tipo *Diamine*, ocorre quando existe um resíduo básico muito próximo de um resíduo básico na superfície da proteína. Nesse caso, assim que as carbonilas de resíduos ácidos são ativadas, elas prontamente sofrem amidação com resíduos básicos vicinais.

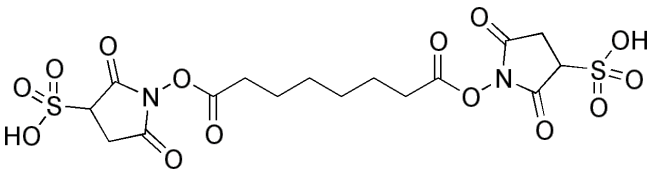

Figura 1.1: Estrutura do suberato de bissulfosuccinimidila, um *cross-linker* utilizado na conexão de grupos que possuam um radical amino lábil. Nota-se, na estrutura, a porção característica do comprimento, o suberato, e também, em ambas as extremidades, as porções responsáveis pela reatividade, no caso, os ésteres.

1.4.2 Obtendo, interpretando e usando dados de XL-MS

A modelagem utilizando dados de XL-MS é uma abordagem relativamente nova na modelagem biomolecular [35] , mas vem ganhando muita notoriedade e, no ano de 2017, junto com a modelagem com dados de SAXS, ganhou uma categoria à parte na competição internacional de proteômica estrutural CASP [22].

Dentre os desafios à viabilidade da técnica de XL-MS para predição de estruturas proteicas está o tratamento e seleção dos dados obtidos [36]. Já existem softwares capazes de processar os múltiplos espectros resultantes dessa análise, atribuindo sinais e assinalando *links* [37, 38, 39], mas a seleção e o *input* de restrições na modelagem ainda obedece a um protocolo artesanal e, muitas vezes, arbitrário, onde é difícil descrever conceitualmente o que é um falso-positivo ou um verdadeiro-positivo. Trabalhos da área reconhecem e lamentam que não

exista ainda um *workflow* automatizado de tratamento e seleção de dados experimentais.

1.5 Propostas desse trabalho

Levando em consideração tudo o que foi exposto nessa breve introdução à situação vigente, desde a exposição do problema da modelagem biomolecular, o problema da amostragem conformacional, a necessidade de confiar nos campos de forças para limitar a busca configuracional, a possibilidade de personalizar os campos de forças para introduzir informação experimental, o advento da técnica de XL-MS e os desafios da modelagem assistida por XLs em fase embrionária, e o vislumbre do problema da seleção dos dados de XLs para injeção no protocolo de modelagem, delimita-se a proposta desse projeto: desenvolver indicadores, métodos, modelos e protocolos baseados em aprendizagem estatística para classificar, qualificar e selecionar dados de XL-MS, que serão usados para auxiliar a simulação de modelos de proteínas.

1.5.1 Detalhamento dos objetivos

Sugerir, desenvolver, aplicar e qualificar indicadores estatísticos baseados em medidas como variância, correlação, diferenciabilidade, uniformidade, aleatoriedade ou outros tipos de medidas, de maneira automatizada. Utilizar os indicadores e modelos desenvolvidos para a seleção automatizada de restrições topológicas. Gerar sistematicamente conjuntos de modelos que reproduzam os objetivos gerais para diferentes proteínas já conhecidas, avaliando a robustez dos métodos. Publicar um pacote de software para disponibilizar os métodos desenvolvidos à sociedade científica.

1.6 Apresentação dos sistemas proteicos estudados

1.6.1 SALBIII, a proteína epóxido hidrolase/ciclase de síntese da salinomicina

O primeiro sistema que será investigado por modelagem assistida é uma proteína denominada SALBIII, uma enzima da classe das hidrolases que ganhou notoriedade no início dos

anos 2010 por seu protagonismo na biossíntese do fármaco Salinomicina [40], que esteve em alta na pesquisa oncológica à época. Sua estrutura cristalográfica foi aferida e publicada (com resolução inferior a 2 Å) no final do ano 2015 por um grupo de colaboradores que incluía pesquisadores do Instituto de Química da UNICAMP [41]. A estrutura cristalográfica da proteína SALBIII é um homodímero, de forma que apenas uma das cadeias, com uma sequência primária de 134 aminoácidos, foi modelada.

Os motivos para seleção dessa proteína incluem não só a experiência anterior de alguns pesquisadores do Instituto com esse sistema, mas também o fato de que sua sequência não possui homólogos suficientes (em número e conservação) nas bases de dados de proteínas para permitir uma modelagem por homologia; o fato de que os dados experimentais já haviam sido coletados para essa proteína e que já havia estudos [42, 43] de sua modelagem, inclusive com *cross-links*; a possibilidade de se utilizar uma biblioteca de fragmentos obtida num momento imediatamente anterior à publicação de sua estrutura cristalográfica e, ao mesmo tempo, contar com a estrutura cristalográfica para avaliar os resultados de cada modelagem.

1.6.2 HSA, a Albumina do Soro Humano

A Albumina do soro humano, codificada como HSA, é uma proteína globular de 585 aminoácidos que responde por mais da metade do teor proteico do soro sanguíneo humano [44]. Essa proteína é composta por três domínios proteicos estruturalmente muito similares [45], denominados HSA-D1, HSA-D2 e HSA-D3. Sua estrutura cristalográfica, com resolução de 2,5 Å, foi depositada no PDB em 1999 [46], resolvida na forma de um dímero.

Diferentemente da SALBIII, a HSA é uma proteína conhecida há muito tempo, vastamente documentada e disponível comercialmente em alta pureza para análises de XL-MS. Isso faz com que ela seja uma complementação prontamente acessível para os testes da metodologia desenvolvida. Os dados de XL-MS da HSA foram, novamente, fornecidos pelo grupo do Prof. Dr. Fabio Gozzo, em colaboração.

Para os experimentos de modelagem, cada domínio da HSA foi tratado como uma problema de modelagem distinto. Essa escolha prática quase sempre é feita, já que a estrutura terciária de domínios globulares costuma ser pouco dependente do estado de oligomerização, sendo

este o caso da albumina. Dados para suporte dessa propriedade podem ser encontrados em [47, 48, 49, 50].

A escolha da HSA se deu porque, a despeito das sequências primárias dos três domínios serem diferentes, os três possuem estruturas terciárias praticamente iguais. Em termos práticos, isso significa que embora os três domínios da Albumina humana sejam estruturalmente similares, os dados obtidos do experimento de XL-MS não necessariamente têm a mesma qualidade para os três, porque se está tratando de três estruturas primárias com reatividade e acessibilidade ao solvente diferentes.

1.7 Estratégia de modelagem selecionada para o Projeto

Para o presente trabalho, tomando por base as experiências anteriores do nosso grupo de pesquisa, o sucesso e notoriedade de cada software em avaliações de escala global e a literatura especializada, bem como as características dos sistemas estudados, decidiu-se empregar uma modelagem *ab initio* utilizando o software Rosetta, mantido e desenvolvido pelo grupo do Dr. David Baker na Universidade de Washington, Seattle. As rotinas desse software serão adaptadas para modelagem assistida utilizando restrições de distância derivadas de XL-MS por meio de personalização do campo de forças REF (Rosetta Energy Function), interno do software.

Capítulo 2

Abordagem e Pontos-Chave

A decisão de empregar o software Rosetta define praticamente todos os critérios estratégicos do projeto, e, por isso, é importante entender a maneira como esse software funciona e de que maneira os dados e algoritmos obtidos e desenvolvidos podem intervir nesse funcionamento e modificá-lo.

2.1 Entendendo o funcionamento do software Rosetta

O software Rosetta é um conjunto modular de aplicações que pode ser compilado e utilizado de diferentes maneiras para múltiplas finalidades. Ele foi inicialmente desenvolvido em 1997 [51] com uma proposta à época inovadora de realizar a modelagem de proteínas usando fragmentos com sequências locais similares à proteína modelada, e em 1999 essa perspectiva do problema de modelagem já obteve resultados significativos no CASP3 [52]. A compilação da versão 3.7 do Rosetta (utilizada nesse trabalho) fornece mais de 150 executáveis à disposição do usuário, mas o mais importante para esse projeto é, sem dúvida, aquele chamado "abinitiorelax". Esse programa contém tudo o que é necessário para executar, do início até o final, uma modelagem multiescala baseada em conhecimento.

2.1.1 Combinando níveis de escala na modelagem

Uma das grandes ideias por trás do Rosetta é combinar diferentes níveis de escala na modelagem. O protocolo *ab initio* utiliza uma representação do tipo grão grosso que se baseia em reduzir a representação de cada aminoácido a apenas sete átomos, a saber:

- O átomo de Oxigênio da Carbonila, O;
- O átomo de Carbono da Carbonila, que participa da ligação peptídica C_{pep};
- O átomo de Carbono alfa à carbonila, C_α;
- O átomo de Carbono beta à carbonila, que faz parte da cadeia lateral, à exceção da Glicina, C_β;
- Um átomo especial chamado $centroid$, cuja função é contabilizar o volume e a carga de todo o resto da cadeia lateral;
- O átomo de Nitrogênio, que participa da ligação peptídica N;
- O átomo de Hidrogênio ligado ao Nitrogênio da Amida, H;

A Figura 2.1 mostra a comparação entre um modelo atomístico e de grão grosso para a proteína Top7 (PDB 1QYS), uma proteína $\alpha\beta$ pequena. Os átomos em verde são os átomos de carbono do *backbone* proteico, enquanto os átomos em ciano são, na figura à esquerda, átomos de carbono das cadeias laterais (que serão reduzidos junto com os demais), e, à direita, átomos do tipo $centroid$.

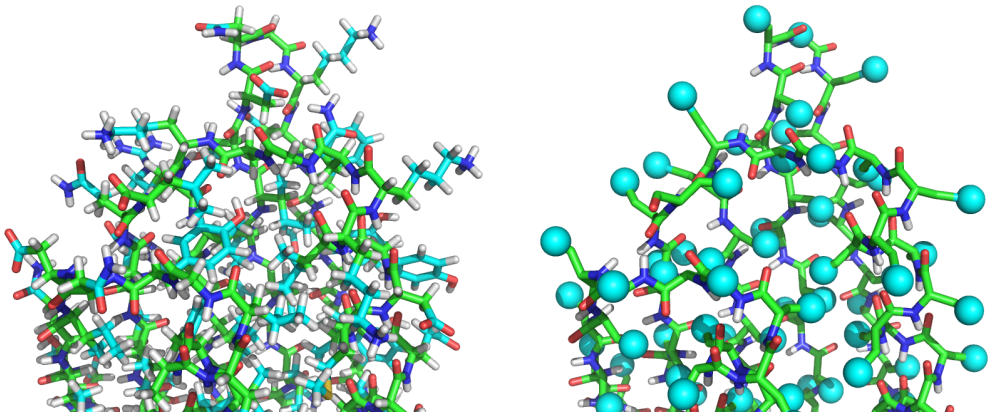

Figura 2.1: Comparação dos modelos tridimensionais atomístico e de grão grosso utilizando a representação centroide do Rosetta para a proteína PDB 1QYS [53]. Reprodução do original do RosettaCommons em [54]

2.1.2 Resumo do protocolo *abinitiorelax*

O protocolo de modelagem *ab initio* do Rosetta realiza sua busca conformacional de forma estocástica por meio de amostragens do tipo Metropolis-Monte Carlo, seguindo aproximadamente o seguinte conjunto de passos para cada modelo que será gerado:

Dada uma sequência primária **P**, uma biblioteca de fragmentos triméricos **B3** e outra biblioteca, de fragmentos nonaméricos **B9**:

Preparação do Modelo

1. Construir uma conformação inicial estendida **M** para o modelo, de grão grosso, considerando apenas o *backbone* proteico;
2. Substituir aleatoriamente trechos nonaméricos de **M** por fragmentos da biblioteca **B9** até que todos os átomos tenham sido substituídos pelo menos uma vez;
3. Calcular a energia inicial **S** (por meio da função *"score"* do próprio software) da conformação estendida, empregando a **Função Energia do Rosetta**, que será discutida posteriormente na subseção 2.2.

A partir desse momento, o software realizará a modelagem propriamente dita, em duas etapas, sendo a etapa inicial uma busca grosseira (*ab initio*) de conformações e a segunda etapa um refinamento (*relax*) do resultado da primeira etapa.

Modelagem *ab initio* (grão grosso)

A modelagem *ab initio* do Rosetta é uma simulação do tipo Monte Carlo com uma representação reduzida (de grão grosso) da proteína. Essa simulação ocorrerá ao longo do que o software denomina "estágios de modelagem", que são conjuntos de milhares de passos MC. O que diferencia cada estágio são os coeficientes (pesos) de cada parcela funcional na Função Energia do Rosetta, a agressividade da substituição de fragmentos e o número de passos MC envolvidos.

Cada singular passo MC envolve a seguinte sequência de eventos:

1. Selecionar um fragmento candidato à substituição, a depender do estágio de modelagem. Esse fragmento será selecionado da biblioteca **B9** em estágios mais agressivos, ou de **B3** em estágios mais brandos;
2. Substituir o fragmento selecionado no modelo **M** gerando um novo modelo **M'**;
3. Calcular a nova energia **S'** de energia do modelo **M'** aplicando novamente a Função Energia do Rosetta;
4. Aceitar a mudança conformacional utilizando para a decisão o critério de Metropolis;

   O critério de Metropolis

   Define-se o critério de Metropolis aquele para o qual a transição entre duas conformações numa simulação Monte Carlo é aceita com probabilidade $0 < p \leq 1$, onde

   $$p = \min\left(1, \exp\left(-\frac{(S' - S)}{T}\right)\right)$$

   Onde S e S' são as energias antes e depois da substituição do fragmento, e T é a temperatura. É importante notar que esse valor de Temperatura é meramente utilizado para ajustar a tolerância da modelagem a mudanças conformacionais ao longo dos ciclos iterativos, onde um valor arbitratriamente alto é empregado no início e sucessivamente diminuído conforme progride a modelagem.

   Esse critério basicamente se traduz da seguinte forma: sempre que a energia do modelo diminuir na substituição, a mudança é aceita. Do contrário, existe uma probabilidade de aceitação dada por $\exp\left(-\frac{(S'-S)}{T}\right)$
5. Caso a mudança seja aceita, o Modelo **M'** substituirá o Modelo **M** e uma nova iteração principiará;
6. Caso a mudança não seja aceita, o modelo **M'** é descartado e o modelo **M** sofrerá uma nova tentativa de substituição de fragmentos, no próximo passo MC (2.1.2).

Na configuração padrão, existem quatro grandes estágios de modelagem *ab initio*, denominados Estágios I, II, III e IV. Nos estágios I e II, são realizados 2.000 passos MC com

amostragem agressiva. No estágio III são realizados 20.000 passos MC e no estágio IV, 12.000 passos MC com amostragem branda são realizados. Portanto, uma modelagem *ab initio* padrão do rosetta é uma simulação com 36.000 passos MC. Esse número pode ser ajustado, a exemplo das modelagens realizadas nesse trabalho, onde o número de passos foi multiplicado por cinco, resultando em simulações com 180.000 passos.

Refinamento *relax* (atomístico)

Após a conclusão da etapa de modelagem grosseira, tem início o refinamento da modelagem. Os átomos centróides do modelo **M** são substituídos por cadeias laterais atomísticas dos aminoácidos e uma base de dados de rotâmeros é utilizada para refinamento das estruturas. Essa base de dados é computada a partir de todas as conformações de um mesmo aminoácido ao longo de todo o PDB, e é atualizada pelo menos uma vez ao ano.

A premissa central do protocolo *relax* não é fazer movimentos bruscos no *backbone* proteico, mas sim explorar finamente o espaço conformacional ao redor do modelo **M**, para encontrar um mínimo que combine um entrelaçamento harmonioso das cadeias laterais e uma relaxação dos ângulos torsionais do *backbone*.

Para isso, será empregado um protocolo muito similar ao descrito para a modelagem *ab initio*, exceto que ao invés da substituição de fragmentos, serão substituídos rotâmeros das cadeias laterais, e a Função Energia do Rosetta será substituída de seus termos centroides para seus termos atomísticos. Nessa substituição, conforme será explicado na seção 2.2, ganham importância os termos relacionados a ligações hidrogênio, é introduzido um termo relacionado à probabilidade *post hoc* de um determinado rotâmero no PDB e os termos repulsivos são melhor descritos.

O protocolo *relax* também é uma simulação MC, porém, diferentemente da modelagem *ab initio*, em que existem quatro grandes estágios de modelagem, no protocolo de relaxação uma estratégia de recozimento é utilizada, onde os coeficientes na Função Energia do Rosetta vão variando lentamente seus valores a cada passo MC. Finalmente, após o refinamento da estrutura, será então gerado um arquivo contendo as coordenadas dos átomos do modelo.

2.2 A Função Energia do Rosetta e os ciclos de modelagem

Todos as energias calculadas em cada iteração da modelagem são fruto da aplicação da chamada Função Energia do Rosetta (em inglês, REF, *"Rosetta Energy Function"*) [55]. Entender esse passo de pontuação dos modelos e o funcionamento dessa função é crucial tanto para compreender o funcionamento do software em si quanto para entender como será possível customizar a energia de cada modelo para levar em consideração os dados de XL-MS que vão assistir a modelagem. A REF é um campo de forças do ponto de vista de que utiliza uma série de funções, parâmetros e variáveis de posição relativa para estimar numericamente um valor de energia potencial para uma molécula.

Nesse projeto, foi empregada uma versão da REF denominada Talaris2014 [56]. Diferentemente dos campos de força clássicos, que possuem apenas alguns poucos tipos de parcelas (ou formatos funcionais) para calcular cada componente do potencial, a Função Energia do Rosetta possui, por padrão, um grande conjunto de termos (mais de 70 deles se aplicam só à modelagem de proteínas) e parâmetros diferentes. A tabela 2.1 lista os termos mais importantes para a modelagem mostrada na seção 2.1.2:

2.3 Agregação dos dados de XL-MS

Depois de compreender como funciona e qual é a composição básica da Função Energia do Rosetta, fica simples entender como é possível utilizar os dados instrumentais para assistir a modelagem. uma vez que:

1. Existe uma forma funcional (um termo, ou parcela) diferente para cada pequena componente relevante da energia de uma molécula,
2. A energia calculada é uma combinação linear de todos os termos ponderados pelos seus coeficientes, e
3. A energia calculada é o critério de aceitação de uma nova conformação na simulação MMC,

Basta acrescentar a essa vasta coleção de termos pré-programados no Rosetta (e, portanto, agregar também à combinação linear que gera a estimativa numérica da energia) um termo novo que leve em consideração os dados de XL-MS [59, 60].

Tabela 2.1: Parcelas da Função Energia do Rosetta, de baixa e alta resolução, tomando por referência a versão Talaris2014 da REF [57, 58].

Termos de baixa resolução (grão grosso, utilizados na fase *ab initio*)	
Parcela	Significado
env	Potencial relativo à hidrofobicidade do resíduo
pair	Potencial relativo a pontes dissulfeto e interações iônicas
vdw	"Van der Waals", parcela energética que contabiliza a repulsão estérica entre resíduos
rg	"Raio de Giração", uma parcela que favorece estruturas compactas (enoveladas) em detrimento de estruturas volumosas (desenoveladas)
cbeta	Potencial que se contrapõe à rg compensando o volume de exclusão decorrente da aproximação de grão grosso das fases iniciais da modelagem
hs_pair	Potencial relativa ao empacotamento entre α-hélices e β-folhas
ss_pair	Potencial relativo ao empacotamento entre diferentes segmentos de β-folhas
hb_bb	*"Hydrogen Bond - Backbone"*, potencial relativo às ligações hidrogênio do *backbone* proteico
Termos de alta resolução (atomísticos, utilizados na fase *relax*)	
fa_atr	Potencial atrativo entre as cadeias laterais de diferentes resíduos
fa_rep	Potencial repulsivo entre as cadeias laterais de diferentes resíduos
fa_sol	Potencial de interaçao atomística com o solvente (solvatação) baseada nas equações de Lazaridis Karplus (solvente implícito)
fa_pair	Potencial de interação eletrostática entre pares de aminoácidos
hbond	Coleção de potenciais relacionados a ligações de hidrogênio em diferentes contextos (*backbone* - *backbone* de curto e longo alcance, *backbone* - cadeia lateral e cadeia lateral - cadeia lateral)
rama	Potencial relacionado aos ângulos diédricos (φ, ψ) do Gráfico Ramachandram
fa_dun	Potencial probabilístico, relacionado à recorrência de um determinado rotâmero da cadeia lateral na base de dados de rotâmeros

Esse termo modificaria a energia de cada modelo, penalizando os modelos que desobedecem às restrições fornecidas. A consequência direta disso é o **aumento da probabilidade do resultado final da modelagem ser uma estrutura tridimensional que obedece às restrições inputadas.**

2.3.1 A informação por trás dos *cross-links*

A fim de propôr um termo adequado para uso dos dados de XL-MS, é crucial entender o tipo de informação que esses dados podem oferecer. Partindo do pressuposto que o *linker* reage única e exclusivamente com a porção superficial da proteína, quando um *cross-link* efetivo (funcionalizado em ambas as extremidades) é observado, a interpretação deve ser que a distância topológica entre

os dois aminoácidos conectados deve ser menor ou igual ao comprimento do linker.

É crucial compreender desse enunciado que os dados de XL-MS podem apenas fornecer uma distância máxima que separa dois aminoácidos, e essa distância dependerá apenas da natureza dos aminoácidos conectados e do *linker* em questão.

Portanto, a forma funcional proposta para o termo de energia associado aos *cross-links* deve ser de tal natureza que não penalize os modelos até que essa distância máxima seja superada. No entanto, uma vez atingido o limite do comprimento do *cross-link*, os modelos devem ser progressivamente penalizados.

2.3.2 Propondo e introduzindo uma função

Existe uma discussão - atualmente em curso - em relação ao melhor tipo de potencial a ser aplicado a partir desse ponto. Alguns autores defendem que a penalização linear com a distância basta, pois não há dados que justifiquem qualquer outra parametrização. Outros sugerem que o potencial deve ser harmônico para agravar a penalização conforme se afasta da distância máxima. Outros, ainda, procuram estabelecer formas funcionais que façam mais sentido, utilizando dinâmica molecular dos linkers em solução e na superfície da proteína, ou abordagens estatísticas. Para esse trabalho, foi selecionada uma forma funcional linear, por acreditar-se ser suficientemente simples para infligir penalização sem traçar hipóteses demais sobre a dinâmica dos *cross-links*. O potencial harmônico foi evitado porque poderia se tornar contraproducente em casos extremos.

A função, portanto, deve ser da seguinte forma, denominada comumente *FLAT-LINEAR*:

$$V(x, V_0, L, a) = \begin{cases} V_0 & \text{para } 0 \leq x \leq L \\ V_0 + a(x - L) & \text{para } x > L \end{cases} \tag{2.3.1}$$

Onde x é variável que representa a distância entre os resíduos no modelo em questão, V_0 é um potencial-base aplicado a todos os pares de resíduos, L é o comprimento máximo do *linker* em questão e a é o coeficiente angular da reta que representa o potencial acima de L. Essa função foi implementada no Rosetta com quatro parâmetros, a saber: V_0, x_0 l e a, onde $x_0 + l = L$. Para customizar a Função Energia do Rosetta, declara-se cada parcela (relativa a cada *cross-link*

observado) num arquivo separado de *input* à aplicação *abinitiorelax*, conforme o código 2.3.2 a seguir

Código 2.1: Exemplo de declaração de restrições no protocolo *abinitiorelax*

```
1 AtomPair  CB    1  CB    8  LINEAR_PENALTY  9  0  9  1
2 AtomPair  CB    4  CB    8  LINEAR_PENALTY  5.15  0  5.15  1
3 AtomPair  CB    4  CB    9  LINEAR_PENALTY  8.35  0  8.35  1
4 AtomPair  CB    6  CB    8  LINEAR_PENALTY  10.9  0  10.9  1
```

No exemplo mostrado, são declaradas quatro parcelas diferentes de mesma forma funcional, cada uma com seus parâmetros, para cinco restrições diferentes. A exemplo da primeira linha, a porção `AtomPair CB 8 CB 1` indica que essa parcela levará em consideração a distância entre o carbono β do resíduo 1 e o carbono β do resíduo 8, e a porção `LINEAR_PENALTY 9 0 9 1` indica que a forma funcional é do tipo *FLAT-LINEAR* com $x_0 = 9$ Å, $V_0 = 0, l = 9$ Å, e $a = 1$

Como assim, "restrição"?

Cabe aqui uma pequena observação: no sentido cru da palavra, aplicar uma restrição, do ponto de vista físico-químico, seria equivalente a fixar uma determinada conformação para um determinado grau de liberdade, reduzindo o número de graus de liberdade a serem amostrados. O que está sendo proposto é diferente: ao personalizar a Função Energia do Rosetta, sem contudo reduzir o número de graus de liberdae, introduz-se um viés para determinadas conformações que obedeçam as restrições, sem impedir a exploração do espaço conformacional.

2.4 Selecionando os dados de XL-MS

2.4.1 Modelagem de Partida

O propósito central da modelagem assistida é obter um conjunto de modelos com qualidade superior à modelagem não-assistida. No escopo desse projeto, deu-se o nome de **modelagem de partida** à modelagem inicial, sem restrições, realizada para cada sistema investigado. Essa modelagem de partida serve, aqui, a dois propósitos: primeiro, ela atua como

referencial em qualquer comparação de incremento de desempenho fornecido pela modelagem assistida e pela seleção de restrições; segundo, viabiliza uma abordagem bayesiana do problema de seleção de restrições, do ponto de vista que gera dados iniciais para a aplicação de qualquer indicador de seleção de restrições que leve em consideração a qualidade dos modelos num passo anterior de modelagem.

2.4.2 Restrições Triviais e seu impacto na modelagem

Um dos pontos centrais desse projeto é que nem todas as restrições dentro do conjunto obtido experimentalmente introduzem a mesma informação (ou o mesmo grau de viés) à modelagem assistida.

Algumas restrições, ainda que sejam consistentes com a estrutura cristalográfica ou com qualquer estrutura que seja considerada de referência para o problema de modelagem, podem ser irrelevantes para o incremento genérico da qualidade da modelagem. Em todos os sistemas estudados, existe um subconjunto de restrições que compartilham entre si determinadas características, a saber:

- Ocorrem em alta frequência em qualquer modelagem (arbitrariamente mais de 80% dos modelos exibem essas restrições), independentemente de terem sido introduzidas explicitamente na modelagem;
- Geralmente são muito pouco espaçadas do ponto de vista da sequência primária, restringindo aminoácidos que já são razoavelmente vicinais (separados por no máximo 10 resíduos e geralmente pertencentes ao mesmo segmento de α-hélice ou β-folha;
- Em alguns casos são muito correlacionadas com outras restrições (geralmente com as mesmas características), ou seja, fornecem em alguns casos informação estrutural repetida;

A essas restrições, no escopo desse trabalho, demos o nome de restrições **triviais**. Nos testes preliminares realizados com a proteína SALBIII foram experimentalmente aferidas 156 restrições das quais apenas 29 são consistentes com a estrutura cristalográfica. No entanto, um exame mais minucioso aponta que 9 dentre as 29 restrições cristalográficas (aproximadamente 30%) exibem as características das restrições triviais. A figura 2.2 permite visualizar

estruturalmente a diferença entre as restrições triviais e não-triviais. Na estrutura à esquerda, foram destacadas, em púrpura, todas as 29 restrições validadas contra a estrutura PDB-5CXO, e à direita, em azul, apenas o subconjunto de restrições triviais. É visualmente perceptível a localidade estrutural e alta redundância dessas restrições.

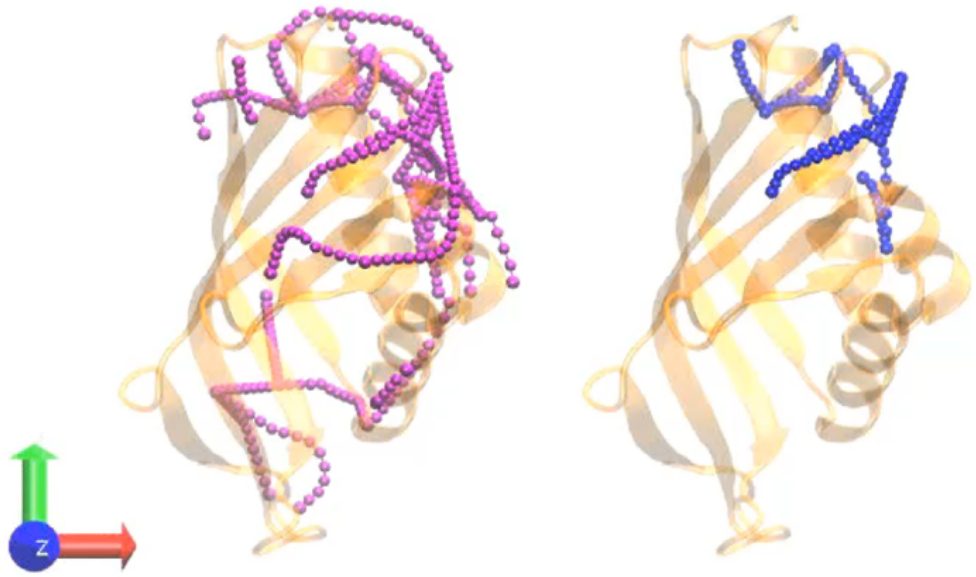

Figura 2.2: Representação das restrições experimentais totais (em púrpura) e triviais (em azul) para a SALBIII, validadas pela estrutura cristalográfica (representada em laranja). As restrições estão representadas como colares de contas que emulam a conformação do *linker* ligado à proteína.

As 9 restrições assinaladas em azul na figura 2.2 são as seguintes: 4CB-8CB, 4CB-9CB, 6CB-8CB, 6CB-9CB, 8CB-9CB, 13CB-17CB, 111CB-113CB, 111CB-115CB, 113CB-116CB. Todas elas obedecem aos critérios de trivialidade apresentados.

Experimentando com as restrições triviais

Para tentar investigar o impacto das restrições triviais na modelagem, em um dos muitos testes realizados ao longo do trabalho foi obtido um determinado conjunto de 37 restrições (denominado "BIS") que não continha nenhuma restrição trivial, e uma modelagem foi realizada usando esse conjunto. Em seguida, todas as 9 restrições triviais foram adicionadas ao conjunto, gerando um novo conjunto denominado "BIS+TRIVIAL", e uma nova modelagem foi realizada. A figura 2.3 mostra uma sobreposição das distribuições de qualidade dos modelos após ambas as modelagens.

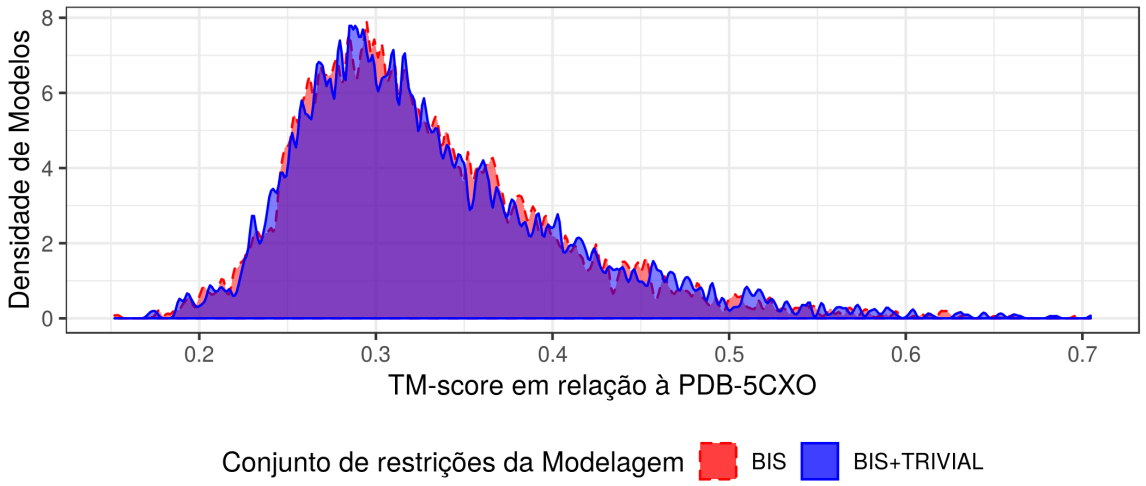

Figura 2.3: Distribuição de qualidade dos modelos obtidos em modelagem comparativa dos conjuntos BIS e BIS+TRIVIAL

O eixo das abscissas dessa distribuição, referente à qualidade de cada modelo, será melhor explicado na seção 2.5.1, mas é importante denotar, por hora, que considera-se um modelo de topologia adequada aquele que alcança um valor superior a 0.5 nessa escala.

A Figura 2.4 a seguir permite comparar ainda melhor as distribuições.

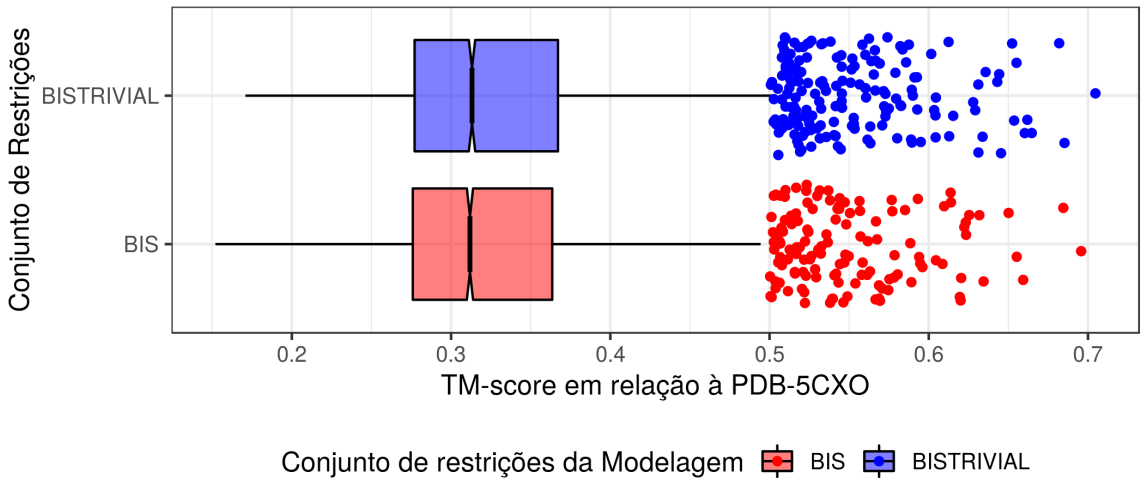

Figura 2.4: *Boxplot* dos modelos gerados em cada uma das modelagens comparadas (BIS e BIS+TRIVIAL). Foi aplicada uma perturbação vertical nos pontos *outliers* para fins unicamente visuais.

Visualmente existe uma grande similaridade não só no formato e na área de ambos os gráficos acima de 0.5 no eixo horizontal, mas também entre as duas distribuições de qualidade dos modelos. Contudo, a fim de confirmar essa hipótese, existem dois experimentos adicionais que

podem ser realizados: um teste de hipóteses e um projeção de similaridade estrutural.

Teste de Hipóteses (Teste U de Mann-Whitney)

Sendo os dados não-paramétricos, sugere-se a aplicação de um teste U de Mann-Whitney, um teste de hipótese não-paramétrico que parte das seguintes prerrogativas: primeiro, que as duas amostras comparadas são indepententes entre si e que os pontos de cada amostra são, também, independentes. Segundo, que as respostas são ordinais. Dentro dessas condições, a hipótese nula do teste de Mann-Whitney é que **as distribuições de ambas as populações são iguais.** O resultado do teste U de Mann-Whitney está expresso na tabela 2.2.

Tabela 2.2: Parâmetros e resultados do teste U de Mann-Whitney

Parâmetro	Valor
U_1	$1{,}23 \times 10^7$
U_2	$1{,}26 \times 10^7$
W	$1{,}23 \times 10^7$
μ_U	$1{,}25 \times 10^7$
σ_U	$1{,}44 \times 10^5$
Z	$\pm 1{,}17$ (bicaudal)
valor p	$0{,}242$

Uma vez computado o valor p, ele deve ser comparado a um valor crítico dado pelo nível de significância do teste de hipóteses, que, por sua vez, é complementar ao nível de confiança. Rejeita-se a a hipótese nula se a probabilidade dela, calculada no teste (valor p), for inferior ao nível de significância proposto. Uma maneira simples de entender o nível de significância é a seguinte: qual é a probabilidade aceita de rejeitar erroneamente a hipótese nula quando ela for verdade?

O valor mais comum empregado em testes bioestatísticos é um nível de confiança de 95%, o que fornece um nível de significância (valor crítico) $\alpha = 0.05$. Uma vez que $0.242 > 0.05$, pode-se dizer que, a partir do resultado do teste U de Mann-Whitney para comparar ambas as populações de modelos geradas, a um nível de confiança de 95%, é impossível descartar a hipótese nula, ou seja, existe uma probabilidade relevante de que as populações pertençam à mesma distribuição.

Dessa observação, é corolário que a adição das 9 restrições triviais a um conjunto de 37 restrições (representando um incremento de 25% no tamanho do conjunto) não foi suficiente para alterar significativamente (a um nível de confiança de 95%) a distribuição de qualidade dos

modelos, ou seja, não introduziu viés suficiente para alterar o formato da distribuição de qualidade da modelagem.

Projeção de Similaridades

Depois de confirmar, por meio de um teste de hipóteses, que a adição das restrições triviais não foi suficiente para alterar significativamente a distribuição de qualidade dos modelos, resta ainda avaliar o possível viés estrutural que elas podem imprimir à modelagem. Para fazer essa avaliação, foram manualmente selecionados os modelos de topologia adequada de ambas as modelagens (BIS e BIS+TRIVIAL), que estão representados na forma de pontos na Figura 2.4. Esses modelos foram misturados e estruturalmente alinhados entre si, produzindo uma matriz de similaridades. Em seguida, por meio de uma técnica de redução de dimensionalidade que será melhor explorada na subseção 2.6.3, foi criada uma projeção bidimensional das dissimilaridades entre os modelos, que pode ser observada na Figura 2.5.

Se houvesse uma segregação estrutural significativa impressa à modelagem por meio da introdução de restrições triviais, seriam observadas na Figura 2.5 duas regiões distintas concentrando, cada uma, pontos de uma das modelagens; no entanto, o que se observa é uma mistura de pontos ao longo de toda a projeção, distribuídos aleatoriamente e sem uma clara preferência espacial de uma ou outra modelagem. Portanto, confirma-se aqui, novamente, que as restrições triviais não introduzem nenhum viés - seja ele significativo ou estrutural - à modelagem.

2.4.3 Lidando com as restrições triviais

Portanto, a partir desse ponto foram traçados os princípios de norteamento do trabalho baseados nessa hipótese, que podem ser sintetizados da seguinte forma:

> Deve ser possível propor um indicador de qualidade para cada restrição que leve em consideração alguma medida de qualidade relacionada a cada modelagem e que possa ser empregado num protocolo de recuperação de restrições que se mostre superior a uma seleção baseada apenas em estatística descritiva. Se for incorporado nesse indicador algum mecanismo capaz de evitar as restrições triviais, o uso do conjunto de restrições

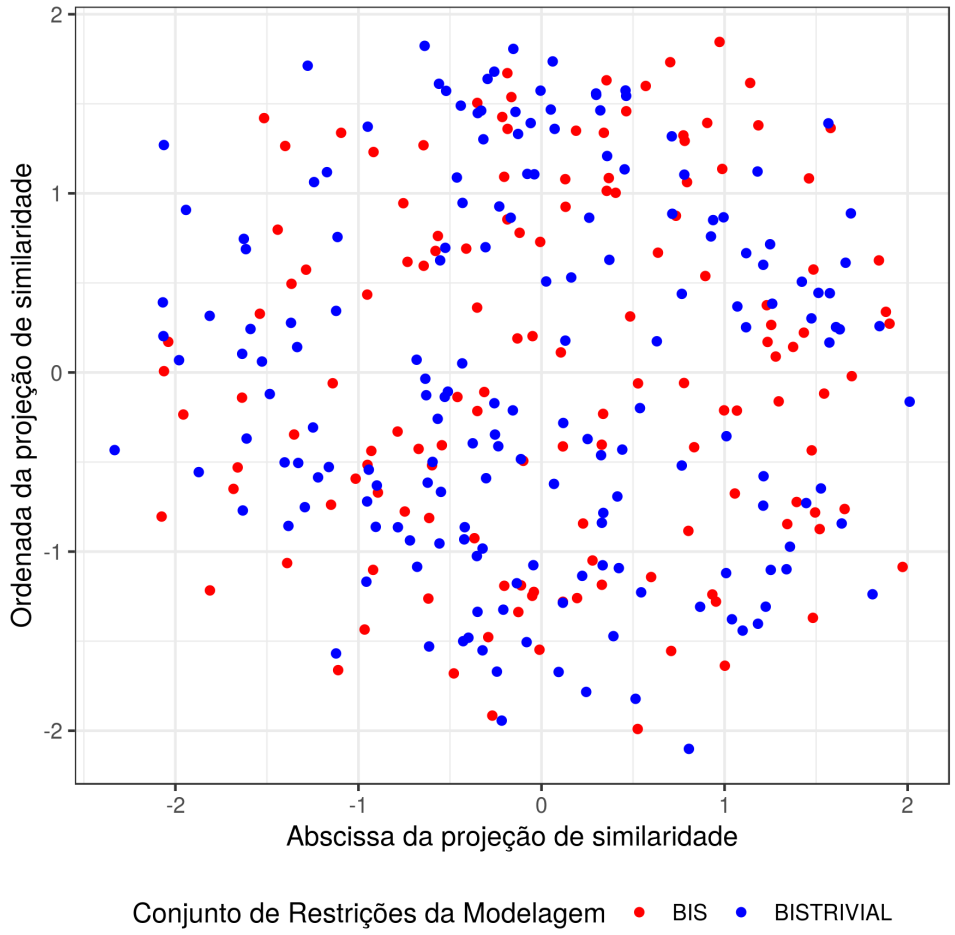

Figura 2.5: Projeção bidimensional das dissimilaridades enrtre os modelos gerados em cada uma das modelagens comparadas (BIS e BIS+TRIVIAL). A distância euclideana entre os pontos é inversamente proporcional à sua similaridade estrutural.

recuperadas pode ser otimizado, tanto do ponto de vista de introduzir restrições que sejam realmente informativas quanto de ampliar a abrangência das restrições para regiões da moléculas que as restrições triviais não cubram.

2.4.4 Coeficiente de correlação ponto-bisserial

Nesse ínterim foi, portanto, proposta a implementação do coeficiente de correlação ponto-bisserial, poderosa ferramenta de avaliação discriminante aplicada originalmente em algumas áreas de estudo da Bioinformática e Psicometria [61] . O coeficiente bisserial é uma medida da correlação entre uma variável fatorial, binária ou dicotomizada (natural ou artificialmente) e uma variável contínua.

Trata-se do caso dicotômico do coeficiente de Pearson, em que uma das variáveis é classificatória ou discriminatória em relação à outra. No contexto da análise linear de discriminantes, o seu significado se torna o de identificar o quão bem uma variável binária - no caso a obediência ou não a cada restrição individualmente – diferencia os modelos segundo um dado critério de qualidade individual.

Seu embasamento matemático prevê que, dado um determinado *score* contínuo S para cada um dos n modelos gerados, e para cada restrição k, cuja frequência pertence ao intervalo $0 > f > 1$, e para a qual cada modelo é classificado como 0 (viola a restrição) ou 1 (obedece à restrição), a correlação bisserial r é dada por

$$r_k = \frac{\overline{S}_1 - \overline{S}_0}{s_n}\sqrt{\frac{n_1 n_0}{n^2}} \tag{2.4.1}$$

Onde $\overline{S}_1$ é o *score* médio para os n_1 modelos que obedeceram à restrição, $\overline{S}_0$ é o *score* médio dos n_0 modelos que violaram a restrição e s_n é a estimativa do desvio-padrão do *score* para todos os modelos, dado por

$$s_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{2.4.2}$$

A proposição desse indicador é um dos principais pontos de inovação deste projeto, sendo extremamente interessante para esse fim porque é capaz de contabilizar a consistência entre cada restrição e a distribuição de qualidade dos modelos, permitindo, por exemplo, identificar as restrições mais correlacionadas com determinada região dessa distribuição e enviesar o próximo passo da modelagem em favor de popular ainda mais essa região de referência.

Há também outros aspectos interessantes, como por exemplo o controle da escala numérica desse valor (que pertencerá sempre ao intervalo $-1 > r_{pb} > 1$), que possui pelo menos um corte não-arbitrário no valor zero [62].

Além disso, e justificando sua escolha, o termo normalizador $\sqrt{\frac{n_1 n_0}{n^2}}$ garante uma penalização algébrica às restrições triviais, para as quais, devido à sua alta frequência, $n_1 \gg n_0$ ou vice-versa, que, no entanto, não inviabiliza a recuperação de restrições que sejam frequentes apenas em modelos de alta qualidade.

2.4.5 Prós e contras dos critérios de seleção

O critério de seleção mais natural, e que foi experimentado anteriormente no grupo [43], é baseado na frequência (recorrência) de uma determinada restrição no conjunto obtido em modelagens preliminares ou anteriores. Esse critério, doravante denominado "frequentista", justifica-se por ser uma medida de proporção - e, portanto, probabilidade - de uma restrição, o que pode levar, indiretamente, a uma interpretação termodinâmica da energia livre de cada restrição. Uma das desvantagens dessa medida, no entanto, é que justamente por sua natureza, ele invariavelmente recuperará primeiro todo o conjunto de restrições triviais para uma determinada modelagem, ocupando lacunas do conjunto recuperado com muitas restrições redundantes e pouco informativas. Baseado nisso, conforme já exposto, o coeficiente de correlação ponto-bisserial é muito interessante, não só por ser uma medida indireta da importância de cada restrição na distinção entre modelos, ou no caso, do viés introduzido por cada restrição na modelagem, mas também pela penalização algébrica das restrições triviais.

Comparando desempenhos e tipos de restrições

Para tentar ilustrar essa propriedade do coeficiente bisserial, apresenta-se, na figura 2.6, um gráfico de dispersão mostrando, em um dos testes preliminares realizados, o coeficiente bisserial e a frequência de cada restrição da SALBIII. Nesse gráfico, os pontos foram coloridos da seguinte forma: em vermelho, restrições experimentais incompatíveis com a estrutura cristalográfica; em azul, restrições experimentais validadas pela estrutura cristalográfica mas consideradas triviais; em verde, restrições cristalográficas e não-triviais. Elipses de confiança foram traçadas utilizando uma distribuição t com índice de confiança de 0.95. Nessa figura, fica perceptível que a recuperação frequentista recupera primeiro as restrições triviais, enquanto que a recuperação por meio do coeficiente bisserial, nesse caso, recuperou apenas uma restrição trivial entre as primeiras 20 restrições recuperadas.

Uma análise hierárquica de agrupamentos foi também realizada para auxiliar a interpretação desse gráfico. Nela, foi utilizado o método da variância mínima de Ward para cômputo dos *clusters*. O resultado dessa análise produziu um dendrograma que está na Figura 2.7.

Nessa figura, estão assinalados alguns grupos para auxiliar a discussão. O *cluster* C1 é

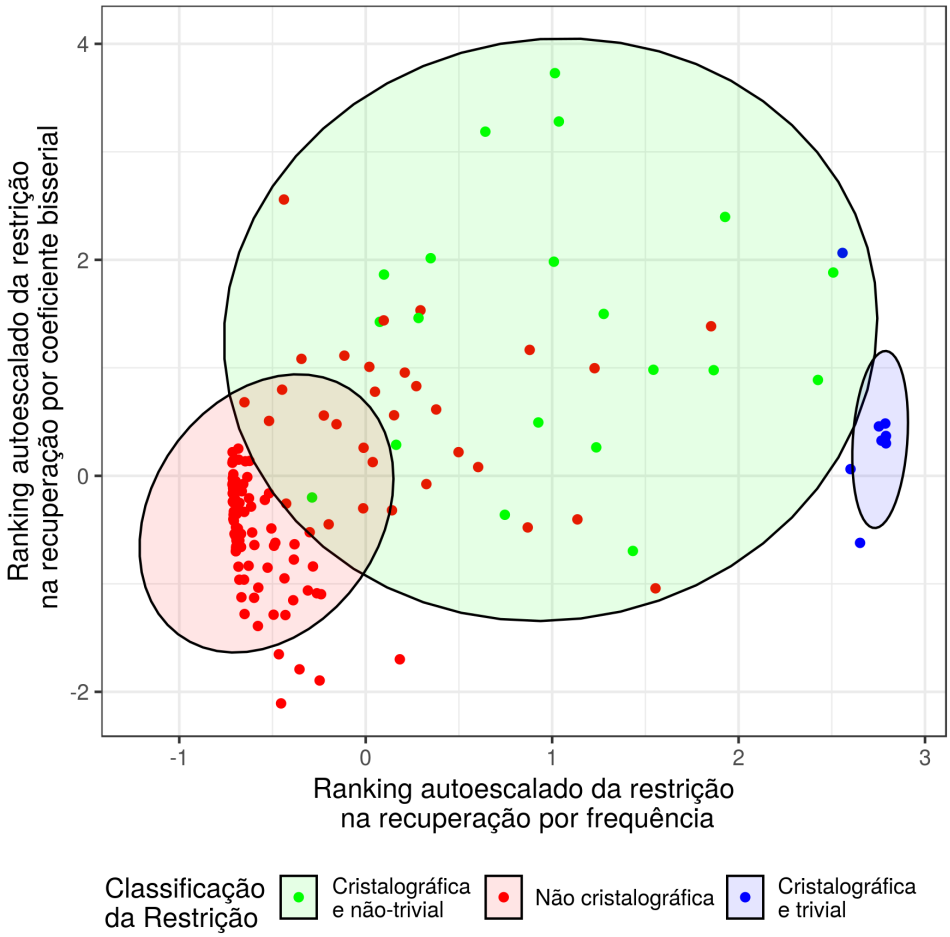

Figura 2.6: Ranking autoescalado das 156 restrições para a SALBIII levando em consideração o coeficiente bisserial (eixo vertical) e a frequência da restrição (eixo horizontal)

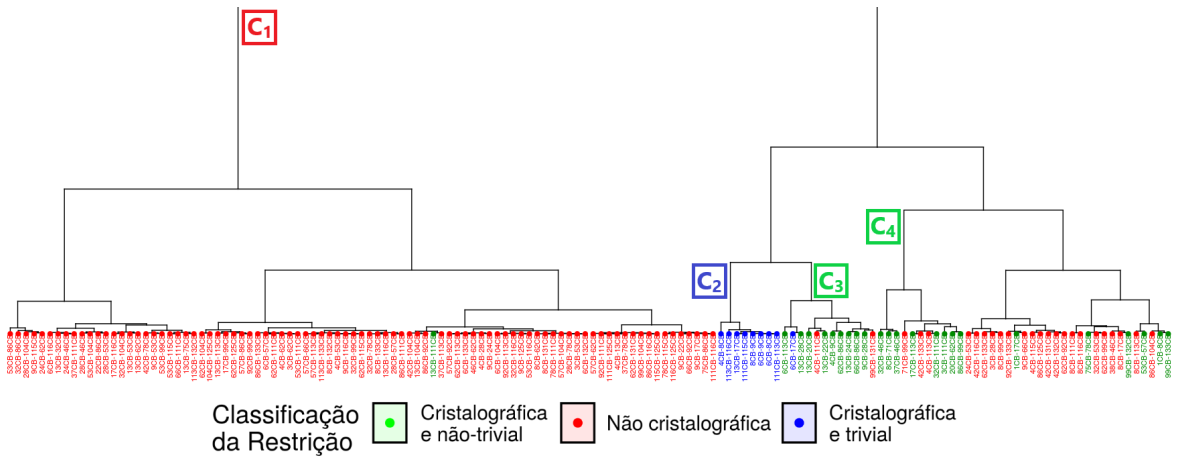

Figura 2.7: Dendrograma para Análise Hierárquica de Agrupamentos realizada nos dados da Figura 2.6

o grupo que, usando qualquer um dos dois critérios (tanto frequentista ou bisserial) nunca é recuperado. Existe apenas um falso negativo nesse grupo, muito infrequente e pouco diferenciador da qualidade dos modelos, o que impede sua recuperação. Em seguida, o *cluster* C2 indica oito restrições (todas elas cristalográficas) que foram agrupadas por terem alta frequência. Essas 8 restrições são consideradas triviais, e são as 8 primeiras recuperadas pelo critério frequentista. O *cluster* C4 apresenta as 11 restrições com maior coeficiente bisserial, das quais 8 são verdadeiros positivos, e todas não-triviais. Já no *cluster* C3 estão as restrições que apresentam valores intermediários para frequência e coeficiente bisserial, sendo recuperadas em qualquer um dos casos.

Se uma análise frequentista recuperar prioritariamente restrições dos *clusters* C2 e C3 e uma análise por coeficiente bisserial recuperar prioritariamente restrições dos *clusters* C3 e C4, é claro que a recuperação frequentista estará recuperando restrições com mais sensibilidade em relação à estrutura cristalográfica. Contudo, 9 das restrições cristalográficas recuperadas na análise frequentista seriam triviais, contra apenas uma restrição trivial na análise por coeficiente bisserial. Se forem considerados verdadeiros positivos apenas as restrições cristalográficas não-triviais, a sensibilidade da recuperação por meio do coeficiente bisserial superaria, portanto, aquela da recuperação frequentista.

O calcanhar de aquiles do coeficiente bisserial

Apesar de todas as vantagens enumeradas sobre o coeficiente bisserial, seu emprego introduz uma complexidade inédita na recuperação de restrições. A equação 2.4.1 requer, claramente, a atribuição de uma determinada pontuação para cada modelagem, simbolizada pelas variáveis $\overline{S}_1$ e $\overline{S}_0$. Logo, foi natural nesse momento sugerir um critério de qualidade preliminar que permitisse qualificar cada modelo gerado individualmente.

2.5 Avaliando a qualidade de uma modelagem

Antes de reportar testes preliminares, é importante entender como se mede o grau de sucesso de uma modelagem nos moldes que foram descritos nesse projeto. Retomando o que foi exposto nas seções introdutórias, a estratégia de modelagem aqui empregada consiste em gerar um grande número de modelos e, desses modelos, selecionar algumas para produzir um ensemble de

estruturas. Embora não pertença ao escopo deste trabalho sugerir a melhor maneira de realizar essa microsseleção, é evidente que, independente de como ela for ocorrer, o sucesso de uma modelagem depende da proporção de modelos gerados com topologia adequada. Quanto maior essa proporção, maior a população relativa de modelos corretos e mais provável a seleção de um modelo adequado. Portanto, fica aqui estabelecido que a qualidade de uma modelagem é uma função da qualidade de cada modelo.

As estratégias de avaliação da topologia de um modelo dependem da quantidade e qualidade de informações sobre o sistema modelado. No caso em que existe uma estrutura de referência disponível, como os casos aqui tratados, os indicadores de qualidade mais adequados são aqueles baseados no alinhamento estrutural entre um modelo gerado e a estrutura de referência, que em todos os casos é a estrutura cristalográfica depositada da proteína em questão. Diferentes medidas de alinhamento estrutural são empregadas em proteômica há muito tempo, e existem muitas maneiras de calculá-las. Os métodos mais famosos são o RMSD (raiz da média quadrática dos desvios, uma medida de dispresão comum na bioinformática), GDT (que computa a proporção de distâncias atômicas dentro de um determinado intervalo predefinido) e o método aqui empregado, denominado Template-Model score.

2.5.1 O *Template-Model score* do alinhamento estrutural

O Template-Model score, ou "TM-score" apresenta uma série de vantagens em relação ao RMSD para o cômputo de alinhamentos estruturais, mas as principais são [63]: seu valor sempre estará entre 0 e 1, independente do comprimento das sequências primárias das proteínas comparadas; além disso, não penaliza modelos de topologia correta, mas que discordam entre si apenas em porções especificamente flexíveis (como as porções N e C terminais, que são naturalmente mais livres na estrutura proteica).

Outro aspecto importante é que o TM-score tem uma identidade probabilística, ou seja, pode ser mapeado para um valor p de teste de hipótese para mesma topologia entre modelos. Um estudo [64] realizado pelos mesmos propositores dessa medida mostra que duas proteínas que apresentam TM-score entre si superior a 0.5 apresentam valores p extremamente baixos para a hipótese de que as topologias comparadas são aleatoriamente diferentes, permitindo portanto dizer

que qualquer modelo com TM-score acima de 0.5 em relação à estrutura cristalográfica provavelmente alcançou a mesma topologia dessa estrutura e, portanto, é um modelo bem-sucedido.

Por isso, durante todo esse projeto, a principal medida de qualidade de estruturas modeladas, e, portanto, de modelagens, empregada é o TM-score do alinhamento de cada modelo em relação à estrutura cristalográfica (PDB-5CXO [41] no caso da SALBIII e PDB-1AO6 [46] no caso da HSA).

2.5.2 Uma breve retratação sobre estruturas cristalográficas

O *Protein Data Bank* é o maior repositório mundial de estruturas macromoleculares, e continha até o final de 2018 150 mil estruturas depositadas, sendo mais de 10 mil só nesse ano [65]. O avanço da proteômica estrutural é, portanto, muito notável, e uma das grandes técnicas responsáveis por resolver e depositar estruturas tridimensionais é a cristalografia de proteínas, que responde por praticamente 90% da base de dados. A primeira estrutura cristalográfica, da mioglobina de esperma de baleia com resolução de 6 $\mathring{A}$ [66], foi determinada em 1958; desde então, dentre todas as técnicas instrumentais já mencionadas para aferição estrutural, a cristalografia é disparadamente a mais predominante, mais antiga e aquela que, até hoje, alcança as melhores resoluções. Não é coincidência que nada menos que 15 prêmios Nobel de química e medicina foram agraciados a cristalógrafos de proteínas [67].

No entanto, é importante retomar brevemente o enunciado aqui proposto (Seção 1.1) para o problema da Modelagem Molecular. A ideia é sempre propôr uma estrutura ou um ensemble de estruturas que represente o estado **nativo** de uma proteína. A estrutura cristalográfica, quando é obtida, geralmente o é a partir de muito esforço e por meio de técnicas avançadas de nucleação e crescimento de cristais. Monocristais ainda são a via de regra, e muita proteínas são cristalizadas com algum tipo de ligante ou cofator em sua estrutura. A própria SALBIII, sistema principal desse trabalho, foi cristalizada na forma de um homodímero com algumas moléculas de hexaetilenoglicol - uma molécula artificial - na sua estrutura [41]. Portanto, acreditar que uma estrutura cristalográfica reflete absolutamente uma conformação nativa em solução fisiológica é ignorar todos esses percalços, e certamente é uma afirmação delicada a se fazer [68].

Contudo, mesmo com todas essas ressalvas, a estrutura cristalográfica ainda é - quando disponível - provavelmente a melhor aproximação do estado nativo ou pelo menos uma conformação provavelmente importante no estado nativo. E, nesse âmbito, é importante concordar com Poincaré, em tradução livre: "É melhor supôr, ainda que admitindo certo grau de incerteza, do que não supôr de forma alguma" [69].

2.6 Explorando critérios de qualidade no coeficiente bisserial

O primeiro critério de qualidade testado com o coeficiente bisserial foi o número total de restrições obedecido por cada modelo. A hipótese era simples: levando em consideração o conjunto total de restrições experimentais, os modelos que obedecessem mais restrições seriam os de melhor qualidade. A fim de testar essa hipótese, por meio desse critério de qualidade, as restrições experimentais foram examinadas e um conjunto das 37 restrições melhor qualificadas dessa forma foi recuperado. Essas 37 restrições foram escolhidas porque eram aquelas para as quais o valor do coeficiente bisserial era positivo, mas nesse momento ainda não havia uma definição do tamanho do conjunto de restrições. Com esse conjunto, alguns testes foram realizados e a figura 2.8 a seguir mostra o resultado da primeira modelagem realizada utilizando esse novo conceito.

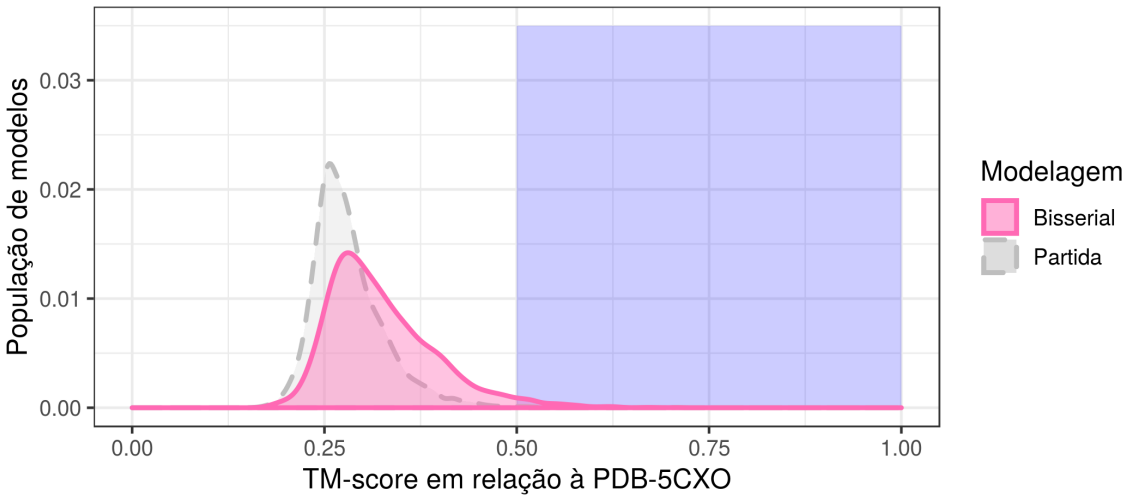

Figura 2.8: Distribuições de qualidade para a modelagem de partida e a primeira modelagem utilizando restrições recuperadas por meio do coeficiente bisserial.

Nessa modelagem, foi observado um incremento de qualidade na recuperação por coeficiente bisserial (em vermelho) em relação à modelagem inicial de referência (em cinza). Além

disso, todas as restrições recuperadas eram do tipo não-trivial, conforme era esperado. O passo seguinte foi, portanto, testar o potencial iterativo desse critério de recuperação de restrições. Foram realizadas 5 iterações, com geração de 5000 modelos em cada uma e recuperação de 40 restrições ao final de cada passo do protocolo. Os resultados estão na Figura 2.9.

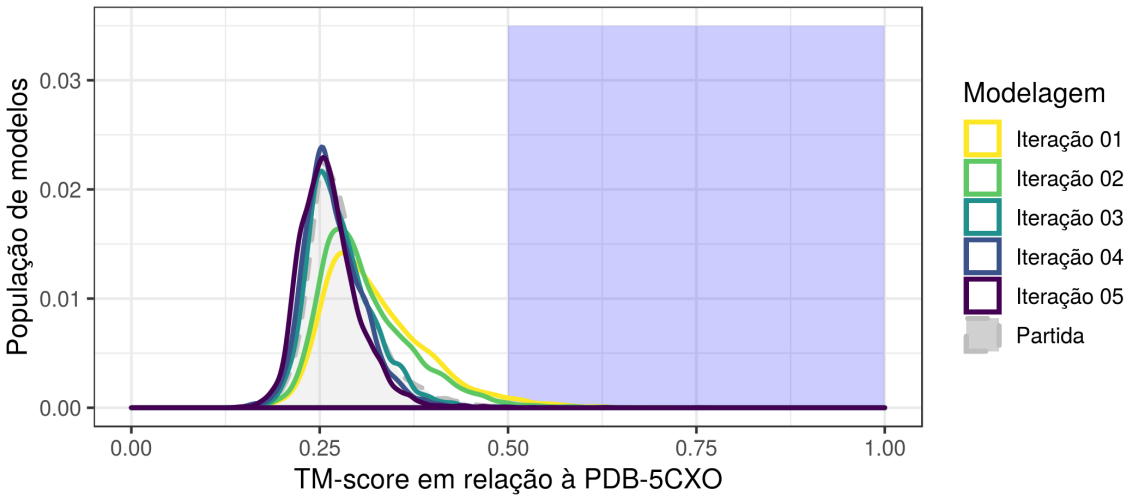

Figura 2.9: Distribuição de qualidade das modelagens (TM-score em relação à PDB-5CXO) em testes preliminares iterativos do coeficiente bisserial em relação ao número total de restrições obedecidas por modelo.

Nesse experimento, o coeficiente bisserial não obteve êxito na recuperação de restrições ou no progresso da modelagem na direção da estrutura nativa. Curiosamente, embora a primeira modelagem seja boa, as iterações causam retrocesso (e não incremento) na qualidade do conjunto modelado. Esse fato, conforme descoberto posteriormente em resultados obtidos paralelamente no grupo de pesquisa, se deve ao fato de que o número total de restrições obedecidas pelos modelos não constitui um bom indicador de sua qualidade. A Figura 2.10 a seguir permite observar que, entre os modelos que obedecem a um alto número de restrições para uma modelagem paralela da SALBIII, existe uma grande dispersão na similaridade em relação à estrutura nativa, invalidando, portanto, o número total de restrições obedecidas como critério de classificação de modelos.

Entretanto, o acompanhamento dos indicadores de desempenho das iterações permite observar que o coeficiente bisserial permaneceu eficaz ao evitar a recuperação de restrições triviais. Em quatro das cinco modelagens realizadas em sequência, não houve **nenhuma** restrição trivial no conjunto recuperado. Conclui-se que o coeficiente de corelação ponto-bisserial obteve sucesso na penalização das restrições triviais, mas o número total de restrições obedecidas por modelo não foi

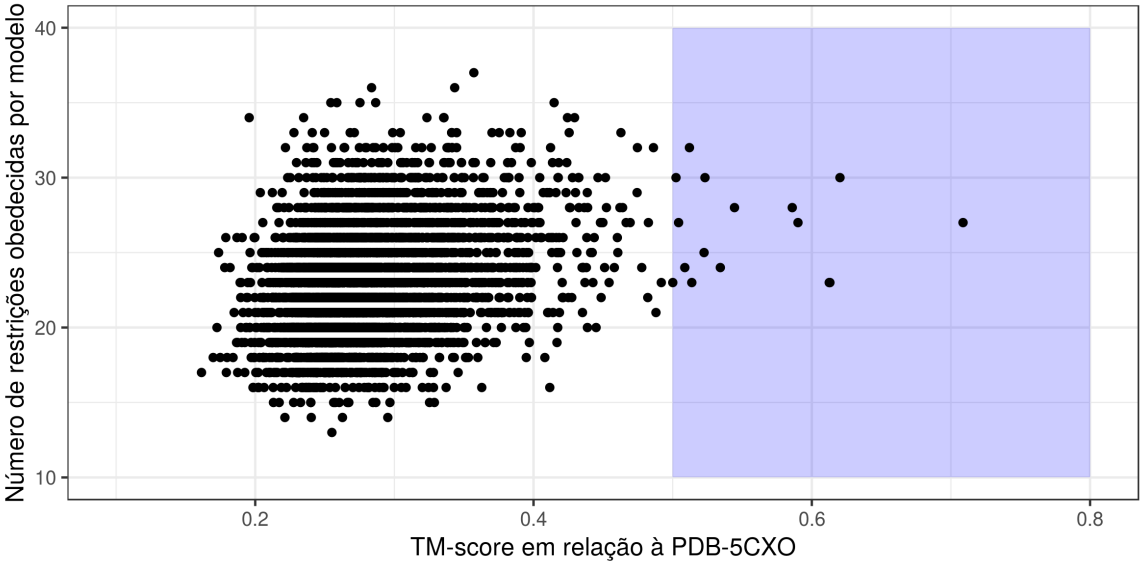

Figura 2.10: Número total de restrições experimentais obedecidas em função da qualidade dos modelos (TM-score em relação à PDB-5CXO) para um dos testes preliminares

uma variável contínua eficaz.

2.6.1 Uma agulha no palheiro: como encontrar um bom modelo?

Assim que se decide empregar o TM-score em relação a uma estrutura de referência como variável contínua, surge imediatemante o questionamento: num caso real, como é que se seleciona uma estrutura de referência?

É importante dizer que propôr uma solução para esse problema foge do escopo desse projeto. Existem muitas maneiras diferentes de fazê-lo, e inclusive na competição CASP existe uma categoria separada só para avaliadores de modelos. No entanto, foi importante, ao longo do projeto, testar e implementar no protocolo desenvolvido pelo menos um tipo de classificador de estruturas proteicas já vigente na comunidade científica.

Casos ideais

Neste trabalho, os quatro diferentes sistemas estudados (SALBIII, HSA-D1, HSA-D2 e HSA-D3) já tinham estruturas cristalográficas depositadas no PDB. É evidente que, para os casos em que isso é verdade, considera-se como caso mais ideal possível aquele em que o mdelo de referência é

sempre a estrutura cristalográfica. Esse caso pode ajudar a estabelecer um tipo de limite superior ao desempenho da técnica, uma vez que se a estrutura cristalográfica é o padrão que mede a qualidade de cada modelagem, provavelmente os melhores resultados serão coletados se ela for justamente a estrutura sempre utilizada como referência.

Além disso, uma pequena flexibilização do caso ideal pode ser também proposta: supondo que exista um classificador de modelos ideal, que sempre seja capaz de encontrar, numa modelagem, a estrutura mais similar à estrutura cristalográfica, com o máximo TM-score em relação a ela - ou seja, **o melhor modelo gerado**. Propõe-se também aqui, num experimento separado, utilizar como referência esse melhor modelo, artificialmente selecionado. Essa pode ser também uma forma de testar se as modelagens seriam capazes de consistentemente alimentar o seu próprio progresso, independentemente do erro introduzido por um classificador de modelos qualquer.

Casos Reais e classificadores de modelos existentes

Quando se pensa na expansão dos protocolos para casos reais, fica clara a necessidade de se implementar algum método de classificação de modelos no protocolo desenvolvido. Existem duas grandes categorias de classificadores de modelos que podem ser empregados nesse tipo de análise. A primeira é denominada "classificadores por consenso", que recebem esse nome porque se baseiam num conjunto de medidas estatísticas que levam em consideração todos os modelos disponíveis numa modelagem.

Um dos classificadores por consenso mais bem-sucedidos nos últimos anos, denominado "Davis Consensus" em referência à Universidade da California em Davis, onde foi desenvolvido, é calculado da seguinte forma: alinham-se todos os modelos um contra o outro, gerando, para n modelos, uma matriz $n \times n$ triangular, contendo $\frac{n(n-1)}{2}$ escores de similaridade. Para cada modelo é, então, calculada a média dos alinhamentos, que constitui o escore individual de cada um.

Essa medida é baseada no fato de que modelos mais centrais no funil de energia livre que representa as diferentes conformações da proteína têm uma densidade de vizinhos – ou seja, estruturas similares a ele – maior, aumentando esse escore. Durante alguns anos, essa medida foi vencedora nas categoria de classificação de modelos da competição internacional CASP (Critical

Assessment of Protein Structure prediction), e ficou muito bem colocada no CASP12 [70], de modo que ele foi um dos classificadores de modelos aqui testados.

A outra grande categoria se refere aos "classificadores de modelo único". Nesse tipo de classificação, um valor numérico representando a qualidade é atribuído a cada modelo individual. Para esse fim podem ser empregados, por exemplo, modelos estatísticos previamente calibrados pelos desenvolvedores, que recebem um conjunto de variáveis que geralmente combina escores de energia livre dos modelos e escores obtidos pelo processamento de bases de dados de proteínas (estes últimos combinando informações como estrutura secundária e superfície acessível ao solvente).

Essas variáveis produzem uma resposta que é computada para cada modelo individualmente. Nesse projeto, foi utilizado um classificador de modelo único chamado ProQ3D, que foi o vencedor da categoria de estimadores de qualidade de modelo único no CASP12 (2017) [71]. Além disso, esse software utiliza alguns binários do próprio Rosetta para realizar sua estimativa, de maneira que a integração com o protocolo de modelagem desde trabalho não demandou a instalação de muitos softwares adicionais. A iniciativa ProQ, que começou há 15 anos [72, 73, 74], já está na sua quarta revisão, na qual *deep learning* foi implementado [75] para melhorar as classificações.

Se um desses indicadores for capaz de apontar um modelo suficientemente bom, seja ele o melhor ou algum modelo muito parecido com o melhor, e, dessa forma, produzir efeitos similares no progresso da modelagem quando comparados à recuperação de restrições baseada na estrutura cristalográfica ou no melhor modelo artificialmente selecionado, teremos obtido sucesso em aplicar uma metodologia **independente da estrutura cristalográfica.**

2.6.2 Testes Preliminares dos classificadores de modelos

Alguns testes preliminares foram realizados com ambos os classificadores de modelos selecionados, o Davis Consensus para classificação por consenso e o ProQ3D no modo de estimativa do TM-score para classificação de modelo único. Na figura 2.11, foram realizadas cinco iterações de modelagem, com 5000 modelos em cada uma e recuperação de 40 restrições a cada passo, onde os critérios de recuperação de restrições foram (A) o coeficiente bisserial utilizando como variável contínua o TM-score de cada modelo em relação a um determinado modelo escolhido por meio do

Davis Consensus e (B) o coeficiente bisserial utilizando como variável contínua o TM-score de cada modelo em relação a um determinado modelo escolhido por meio do ProQ3D-tmscore.

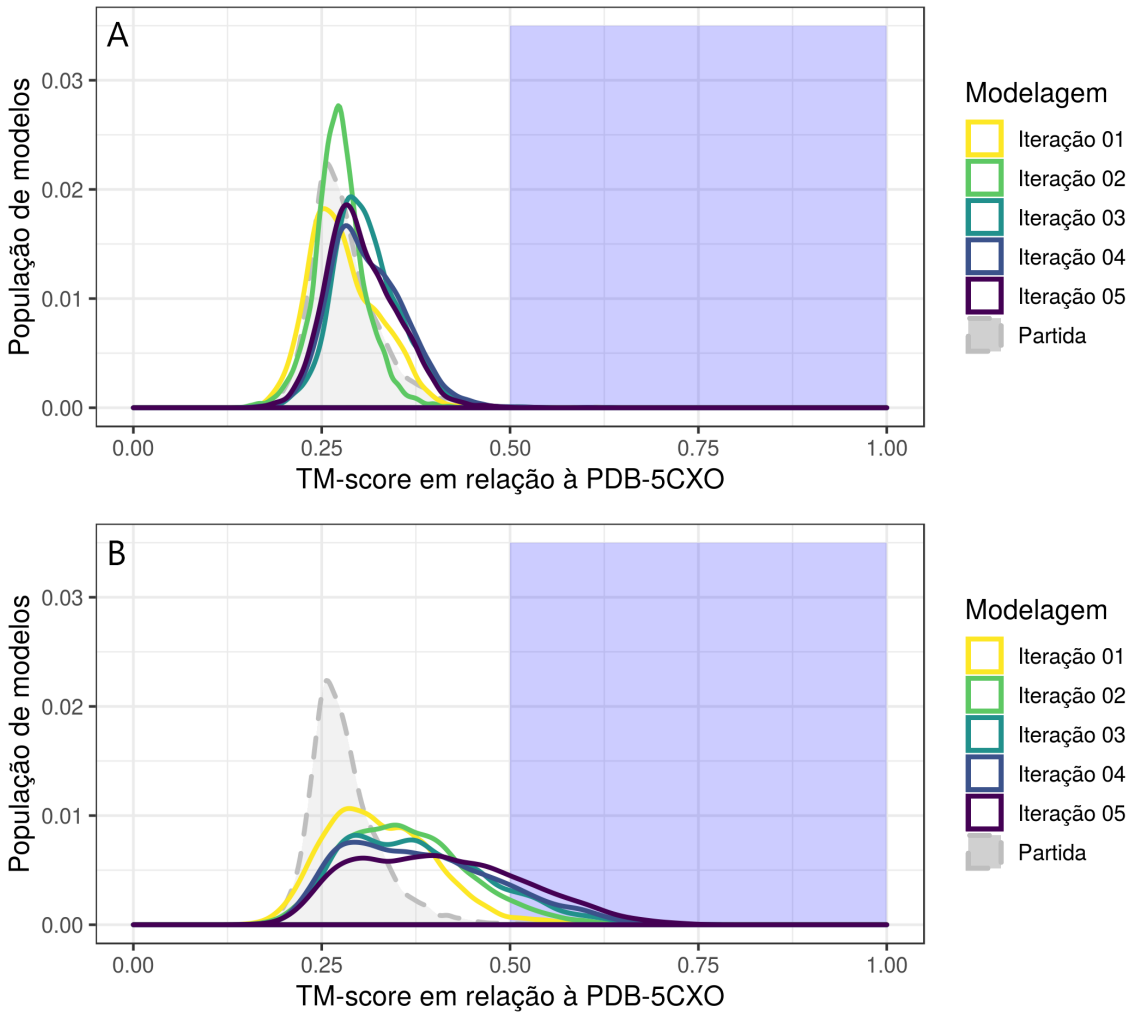

Figura 2.11: Distribuição de qualidade das modelagens (TM-score em relação à PDB-5CXO) em testes preliminares iterativos do coeficiente bisserial usando como variável contínua (A) o TM-score em relação ao melhor modelo selecionado por consenso e (B) o TM-score em relação ao melhor modelo selecionado por ProQ3D-tmscore

Conforme é possível observar na figura 2.11, em (A), o Davis Consensus não foi capaz de recuperar um modelo suficientemente bom a ponto de gerar um viés significativo nas modelagens em direção à estrutura cristalográfica. Um efeito diferente é observado em (B), onde as modelagens são consistentemente melhoradas ao longo das iterações.

Para explorar esse efeito, os modelos selecionados por cada classificador foram comparados individualmente à estrutura cristalográfica. Nesse experimento, percebeu-se que a correlação entre a similaridade de cada modelo com a estrutura cristalográfica e a similaridade com

o modelo selecionado é determinante para o progresso das modelagens, o que é esperado. Percebeu-se também que, no caso dos modelos selecionados por consenso, essa correlação nunca alcançava coeficientes de Pearson superiores a 0.7, enquanto os modelos selecionados por meio do ProQ3D apresentavam comportamento oposto. A Figura 2.12 mostra um exemplo de cada caso. Percebe-se que, em 2.12a, além da pouca correlação linear ($r = 0{,}4$) parece inclusive existir uma distribuição polimodal dos dados, enquanto que, em 2.12b, a correlação linear é muito superior ($r = 0{,}8$). Percebeu-se também que existe uma forte correlação entre o coeficiente de Pearson nesse gráfico e a similaridade do modelo selecionado em relação à estrutura cristalográfica.

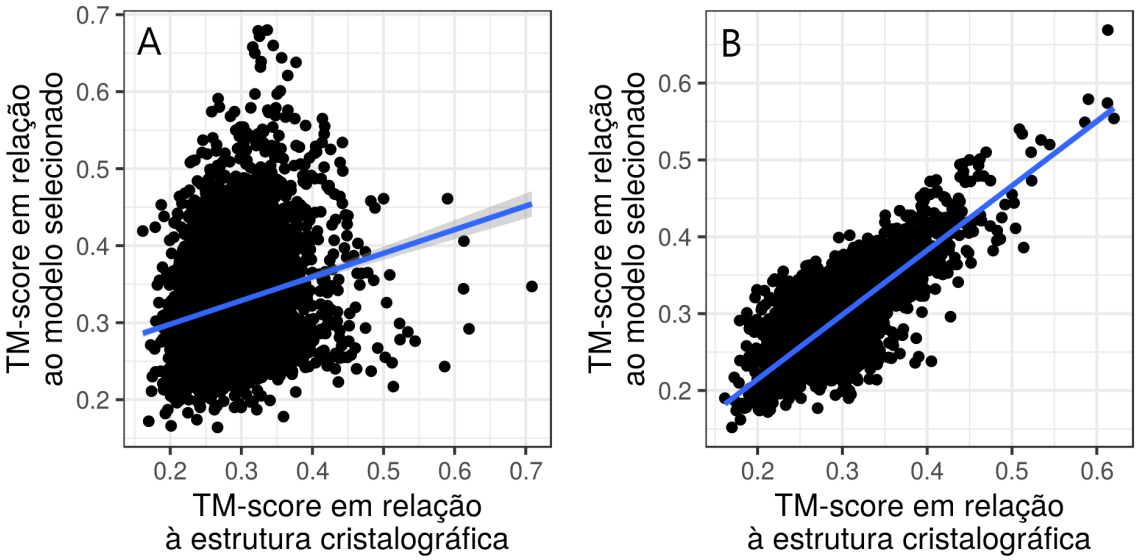

Figura 2.12: Correlação entre as similaridades em relação ao modelo selecionado e à estrutura cristalográfica, para modelos selecionados por consenso (A) e por classificador independente (B). Para efeitos de visualização, foi traçado, em azul, o resultado de uma regressão linear dos dados.

2.6.3 Visualizando o espaço conformacional

Numa tentativa de explorar melhor o espaço conformacional resultante da modelagem, realizou-se um experimento de visualização das conformações relativas entre os modelos gerados. Nesse experimento, foi empregada uma técnica de redução de dimensionalidade desenvolvida por pesquisadores brasileiros denominada **Force-Scheme** [76], que foi escolhida por já ter sido anteriormente mencionada em trabalhos da área de metodologias em modelagem e também por ter sido concebida justamente para preservar, na projeção, as relações de vizinhança entre os pontos. [77]

Para a sua execução, é necessário computar uma matriz de dissimilaridades (distâncias) entre os modelos. Para tal, sugeriu-se calcular esse valor como o inverso do TM-score do alinhamento entre eles ($d_{i,j} = \frac{1}{S_{i,j}}$), de maneira que, sendo $A^{m \times m}$ a matriz de similaridades computada no passo 4 do protocolo de modelagem, cuja diagonal é 1, define-se a matriz de dissimilaridades B^{mxm}, cuja diagonal também é 1, tal que

$$b_{i,j} = \frac{1}{a_{i,j}} = \frac{1}{S_{i,j}} \forall i, j \in [1, m] \tag{2.6.1}$$

A partir dessa matriz B, é calculada a projeção dos pontos pelo seguinte algoritmo:

- Para cada ponto x' da projeção:
  - Para cada ponto $q' \neq x'$ da projeção, computar a direção e o sentido de um vetor $\vec{v}_{x'q'}$
  - Perturbar cada ponto na direção e sentido da resultante dos vetores sobre o mesmo, cada um de módulo δ, dado por:

$$\delta = \frac{d_{(x,q)} - d_{min}}{d_{max} - d_{min}} = \frac{S_{xq}^{-1} - S_{min}^{-1}}{S_{max}^{-1} - S_{min}^{-1}} \tag{2.6.2}$$

O protocolo é repetido iterativamente, até que alcance um determinado critério de convergência (dado por uma perturbação global mínima na projeção) ou alcance um número limite de passos de simulação. O resultado é uma coletânea de pontos cuja distância euclidiana no plano projetado é, de forma geral, proporcional à distância na matriz de dissimilaridades original.

Oliveira *et al.* [77] sugerem que essa figura, computada a partir de dissimilaridades estruturais dos modelos, pode ser interpretada como uma projeção azimutal (planificação transversal) do funil de energia livre de uma proteína. Dessa maneira, espera-se que os modelos de melhor qualidade ocupem posições aproximadamente centrais, e que os pontos sejam distribuídos circularmente, decrescendo de qualidade conforme aumenta a distância entre cada modelo e o centrômero.

A figura 2.13 exibe o resultado dessa projeção para a modelagem inicial não-restringida. Analisando a projeção (B), Percebe-se a disposição circular dos modelos e uma tendência de aumento de TM-score em relação à estrutura cristalográfica em regiões mais centrais. A mesma projeção

foi colorida, em (C), utilizando a densidade de pontos na vizinhança de cada modelo, computada a partir do gradeamento do espaço do gráfico. Essa figura revela a possível causa do fracasso do classificador baseado em consenso: esse tipo de medida tende a selecionar modelos com alta densidade de vizinhos similares. No entanto, a densidade da projeção está altamente dispersa, consequência principalmente da liberdade conformacional conferida à modelagem pela ausência de vieses introduzidos por restrições. Trata-se de uma modelagem não-convergida, onde não se observa de forma alguma a formação de um máximo de densidade central, mas sim de múltiplos máximos em regiões diferentes, que significam conformações diferentes.

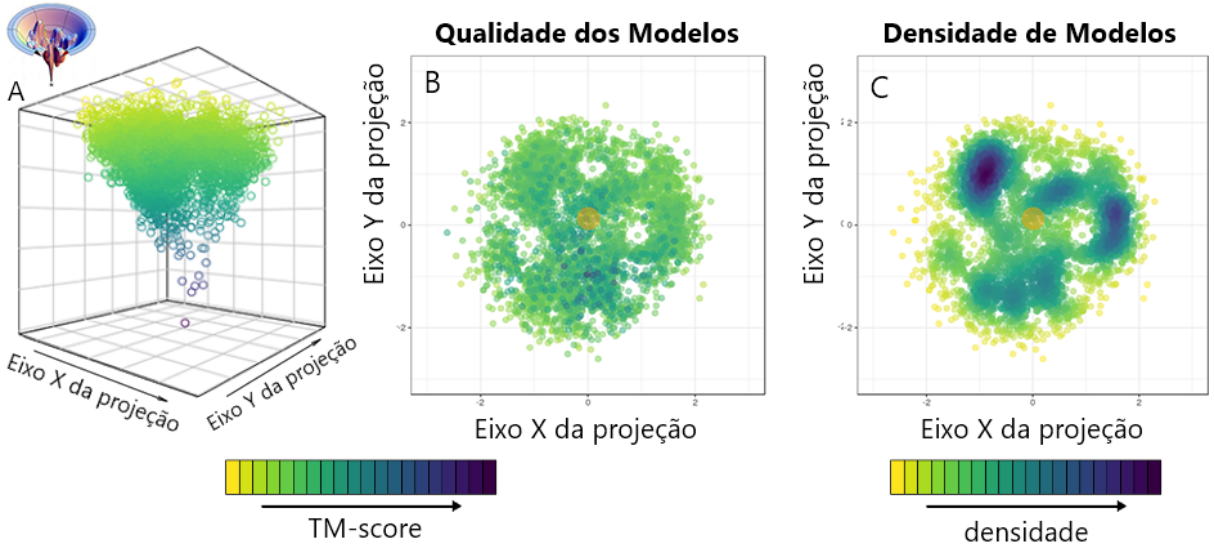

Figura 2.13: Projeção ForceScheme das conformações relativas entre os modelos realizada para os 5000 modelos da modelagem de partida sem restrições da proteína SALBIII

O círculo laranja, que denota a posição aproximada do centrômero de densidades dos modelos, ocupa uma região incêntrica entre três zonas de alta densidade altamente dissimilares. Dessa maneira, a seleção por consenso, ao não encontrar uma região de convergência para a recuperação de modelos, e submetida a uma modelagem em que 99% dos modelos tem TM-score menor que 0.5, acaba por selecionar um modelo de referência distante demais da estrutura cristalográfica para permitir o avanço iterativo.

Em contraste, ao se observar a figura 2.14, percebe-se uma situação de convergência, à medida que se observa um único *cluster* de alta densidade (C), que concentra também os melhores modelos (B). Nesse tipo de modelagem, que foi obtida a partir de cinco iterações do recuperador bisserial utilizando o modelo selcionado pelo ProQ3D, já é possível aplicar medidas de consenso e

outras estatísticas descritivas de tendência.

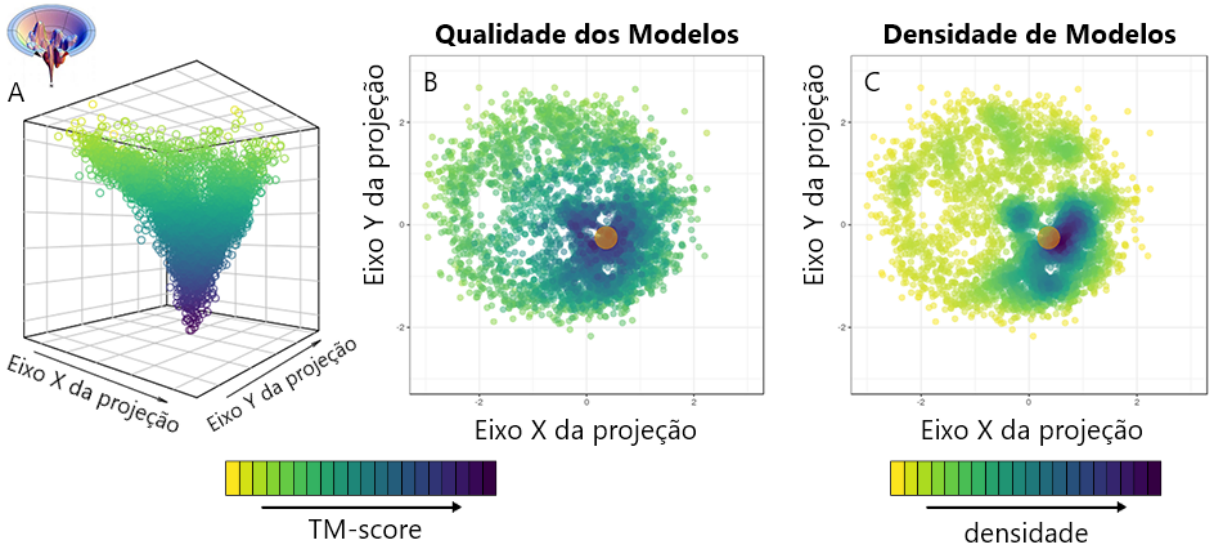

Figura 2.14: Projeção ForceScheme das conformações relativas entre os modelos realizada para 5000 modelos de uma modelagem convergida da SALBIII

Capítulo 3

Metodologia Desenvolvida

3.1 Concepção

Conforme já mencionado, esse projeto se presta a propor métodos e indicadores estatísticos que, por meio da combinação entre os dados experimentais obtidos e análise estatística de modelagens sucessivas, permitam recuperar analiticamente um conjunto de restrições de confiabilidade competitiva, colaborando para a eliminação de arbitrariedades e, portanto, para o robustecimento e enriquecimento da combinação entre a técnica de XL-MS e as modelagens computacionais. Além disso, se propõe também a propôr um procedimento padrão de modelagem que empregue o software Rosetta e os diversos softwares desenvolvidos pelo nosso grupo de pesquisa, além dos softwares de terceiros elencados no capítulo de abordagem, de maneira integrada.

A metodologia proposta é dividida em duas partes principais. Na primeira delas, é realizada uma única vez para cada sistema de interesse, e se refere à coleta de informações sobre o sistema e geração dos arquivos de entrada sempre necessários aos softwares utilizados. A segunda fase, se refere ao processo iterativo de modelagem, seleção e aplicação de critérios de recuperação de restrições.

3.2 Coleta de dados e preparação da modelagem

1. Obtenção da sequência primária do sistema de interesse no formato FASTA.

   A sequência primária é o identificador da proteína, e corresponde à sequência de aminoácidos

na sua estrutura. O formato FASTA foi desenvolvido por biólogos computacionais para simplificar a codificação dessa informação: ele consiste de um vetor de caracteres onde cada letra corresponde a um aminoácido.

2. Obtenção da estrutura cristalográfica da proteína, quando existente, por meio de busca no RSC PDB (Protein Data Bank) [65].

   A estrutura cristalográfica é utilizada como referência para avaliar a qualidade dos modelos obtidos a cada iteração.

3. Busca sequencial de fragmentos proteicos em bases de dados de estruturas de proteínas por meio do software Robetta, para construção da biblioteca de fragmentos (B3 e B9).

   Conforme já mencionado, o software de modelagem Rosetta, empregado neste projeto, requer para seu funcionamento a disponibilização de uma biblioteca de fragmentos para estimar as estruturas secundárias, construir os modelos e permitir a exploração do espaço conformacional. O software Robetta, disponível em servidor público, permite buscar automaticamente esses fragmentos e os exporta já no formato adequado para leitura pelo Rosetta.

4. Obtenção da lista de restrições identificadas experimentalmente por meio de XL-MS.

   Os dados de XL-MS foram cedidos, em colaboração, pelo grupo Dalton MS Group, sob coordenação do Prof. Dr. Fábio Gozzo. Detalhes pode ser encontrados em [34].

5. Geração do arquivo de input do software TOPOLINK a partir da descrição dos experimentos de XL-MS.

   O software TOPOLINK [78] é um pacote de funções desenvolvido para computar distâncias topológicas entre átomos na superfície de proteínas e validar modelos estruturais com base em dados de XL-MS. No arquivo de *input* de cada análise, é importante declarar os tipos de *cross-links* possíveis e suas distâncias máximas consideradas e também os *links* observados. Um exemplo da sintaxe utilizada nesse passo está no código 3.1

Código 3.1: Exemplo de declaração de restrições possíveis e observadas no arquivo de *input* do TOPOLINK

```
1  experiment XL
2
3  linktype   GLU   all   all   CB   GLU   all   all   CB   16.7
4  linktype   GLU   all   all   CB   ASP   all   all   CB   15.4
5  linktype   ASP   all   all   CB   ASP   all   all   CB   14.1
6  linktype   ASP   all   all   CB   LYS   all   all   CB   9.6
7  linktype   GLU   all   all   CB   LYS   all   all   CB   10.3
8  linktype   GLU   all   all   CB   SER   all   all   CB   7.1
9  linktype   ASP   all   all   CB   SER   all   all   CB   6.2
10 linktype   LYS   all   all   CB   LYS   all   all   CB   21.8
11 linktype   LYS   all   all   CB   SER   all   all   CB   18.0
12 linktype   SER   all   all   CB   SER   all   all   CB   14.1
13
14 observed ASP A 35 ASP A 54
15 observed ASP A 35 ASP A 57
16 observed ASP A 35 ASP A 67
17 observed ASP A 35 LYS A 41
18 observed ASP A 35 SER A 127
19 observed ASP A 47 GLU A 119
20 observed ASP A 47 LYS A 149
21 observed ASP A 53 ASP A 54
22 observed ASP A 53 ASP A 57
23 observed ASP A 54 ASP A 64
24 observed ASP A 57 ASP A 64
25
26 end experiment XL
```

No exemplo mostrado no código 3.1, a linha 1 inicia o bloco em que serão declarados os parâmetros experimentais; nas linhas 3 a 12, são declarados todos os tipos de *linker*, levando em consideração o tipo de aminoácido de cada extremidade, o átomo de referência e, principalmente, o tamanho considerado do *linker*. Para esse trabalho, as distâncias consideradas estão na tabela 4.2. Em

seguida, nas linhas 14 a 28, são declarados os links observados. Na linha 14, por exemplo, declara-se que foi observado um *link* entre o a Aspartato de posição 35 na Cadeia A e o Aspartato na posição 47 também na Cadeia A.

6. Determinação das condições de modelagem num arquivo de opções do software Rosetta

   As condições de modelagem são pequenas decisões que precisam ser tomadas em relação ao funcionamento do rosetta, como, por exemplo:

   (a) Número de ciclos dos protocolos *ab initio* e *relax*;

   (b) Número de modelos gerados por iteração (n);

   (c) Formato de gravação dos modelos gerados;

   (d) Termos personalizados na REF;

   (e) Pesos e coeficientes da parte personalizada da REF.

7. Modelagem de partida

   Uma modelagem inicial, sem nenhuma restrição, é realizada para cada sequência primária. Nessa modelagem, geralmente é gerado um número de modelos superior àquele de cada passo do protocolo iterativo.

8. Cômputo opcional de quaisquer variáveis necessárias ao classificador de modelos escolhid.

   Neste trabalho, empregou-se o software ProQ3D para estimar uma qualidade demoninada ProQ3D-tmscore para cada modelo avaliado. Esse score é uma previsão realizda por um modelo calibrado pelos desenvolvedores do ProQ3D, que recebe como *input* dois conjuntos de variáveis: um conjunto denominado **perfil**, que é sempre o mesmo para a mesma sequência primária, e outro conjunto de energias calculadas para cada modelo. Portanto, é importante, ao preparar a modelagem de qualquer proteína, estabelecer esse perfil para uso do ProQ3D. Nele, estão incluídos [74]:

   (a) Estrutura secundária estimada da proteína, por meio da execução do software BLAST [79] na base de dados UNIREF90 [80], com consequente processamento dos dados obtidos por meio do software PSIPRED [81];

   (b) Superfície acessível ao solvente estimada da proteína, por meio do software SSPRO4 [82];

   (c) Informação de Conservação de sequência, por meio da execução do software BLAST na base de dados UNIREF90;

3.2.1 Iterações das modelagens

1. Execução do protocolo de modelagem do Rosetta usando como base as opções definidas no arquivo de condições de modelagem

   A modelagem do rosetta é trivialmente paralelizável, e se o software for compilado com a opção de utilizar uma biblioteca MPI, a modelagem inteira pode ser executada com um único comando.

2. Extração de arquivos .pdb individuais para cada modelo gerado no protocolo de modelagem

   Para diminuir o tamanho dos arquivos criados durante a modelagem, os modelos de proteínas são gravados em disco no formato *silent*; no entanto, o formato padrão para arquivos de estruturas tridimensionais é o formato .pdb, de modo que é necessário extrair os arquivos no formato adequado.

3. Execução paralelizada do software TOPOLINK em todos os modelos gerados, a partir dos arquivos de input produzidos na preparação da modelagem

   O software TOPOLINK [78] irá validar as restrições experimentais em cada uma das estruturas geradas, produzindo um *log* que dirá, para cada modelo, se cada restrição foi obedecida ou não.

4. Execução paralelizada do software LOVOALIGN para alinhamento de todas as n estruturas geradas, produzindo um total de $(n-1)$ arquivos de alinhamento com um total de $\frac{(n-1)(n-2)}{2}$ alinhamentos realizados

   O software LOVOALIGN [83] é um pacote de funções de alinhamento estrutura de proteínas que usa otimização do menor valor ordenado (LOVO) para buscar a condição de máximo alinhamento e computar os escores correspondentes.

5. Compactação dos $\frac{(n-1)(n-2)}{2}$ arquivos de alinhamento obtidos

6. Execução do software G-SCORE para cômputo dos scores individuais de centralidade de cada modelo, com cortes que variam desde 0.40 a 0.85 para cada modelo.

   O software G-SCORE [84] é um pacote de funções desenvolvido pelo nosso grupo de pesquisa, que gera escores de consenso para os modelos. Esses escores são baseados na contagem de

modelos numa determinada análise de agrupamento realizada por meio do TM-Score intra-modelagem. Foi elaborado para ter uma tradução termodinâmica da energia livre de um determinado modelo com base no número de vizinhos. O G-SCORE foi implementado e calculado nesse protocolo para uso futuro e potencial desenvolvimento de sua aplicação.

7. Execução única do software LOVOALIGN para alinhamento dos n modelos em relação à estrutura cristalográfica (Quando disponível)
8. Análise e tratamento automatizados dos dados e arquivos de output por meio das rotinas do software ZedXL desenvolvido durante esse projeto, com recuperação de um conjunto de restrições para um determinado critério e posterior criação ou Modificação do arquivo para customizar a Função Energia do Rosetta
9. Substituição do arquivo de customização da Função Energia do Rosetta e início de uma nova iteração.

3.3 O software ZedXL

O software ZedXL é o principal produto desse projeto. Ele foi desenvolvido como uma biblioteca para a linguagem livre R [85], e será, ao final do projeto, documentado e disponibilizado em código aberto a toda a comunidade.

O pacote combina quatro grandes categorias de funções: rotinas para ler, escrever e processar os diferentes tipos de arquivo gerados ao longo desse protocolo; algoritmos numéricos para cálculo de diferentes figuras estatísticas de cada restrição, com destaque para o coeficiente de correlação ponto-bisserial; agrupamentos de funções para análise de restrições baseados nas figuras estatísticas, incluindo estatística descritiva básica, análise exploratória de dados, análises de agrupamentos e métodos de redução de dimensionalidade; rotinas de geração de gráficos e figuras.

A função do ZedXL é simples. Ele receberá como principal entrada um conjunto de arquivos de log do TOPOLINK validando cada restrição experimental em cada modelo. Além disso, receberá um conjunto de arquivos de alinhamento e classificação para cada modelo, como os arquivos de saída do ProQ3D e G-SCORE, os logs do LOVOALIGN para alinhamento intramodelagem e, quando houver uma estrutura cristalográfica disponível, o log do LOVOALIGN para os alinhamentos

com a mesma. O pacote de software processará então esses dados, a partir dos diferentes critérios de recuperação de restrição já implementados, fornecendo como principal saída um arquivo no mesmo formato daquele mostrado no código 2.3.2 para customização da Função Energia do Rosetta.

Uma pequena sequência dos passos seguidos pela principal função do pacote, preparada para execução automatizada em scripts *bash*:

1. Leitura e processamento de todos os n arquivos de log do TOPOLINK;
2. Leitura e processamento dos $\frac{(n-1)(n-2)}{2}$ arquivos de alinhamento intra-modelagem;
3. Opcionalmente, leitura e processamento de todos os arquivos de output do G-SCORE e ProQ3D;
4. Quando disponível, leitura e processamento do arquivo de alinhamento contra a estrutura cristalográfica;
5. Cômputo da matriz binária de restrições para a modelagem;
6. Cômputo da matriz de escores de modelo e cálculo de escores complementares;
7. Cálculo dos escores das restrições;
8. Ordenamento das restrições segundo o critério de recuperação escolhido;
9. Geração e gravação do arquivo de entrada do Rosetta, criado automaticamente nos padrões do item 2.3.2.

Capítulo 4

Testes da Metodologia

4.1 Dados da modelagem

4.1.1 Dados experimentais de XL-MS

Como parte do preparo das modelagens, foram recebidos os dados de XL-MS de nossos colaboradores no Dalton MS Group, do Instituto de Química, UNICAMP. A tabela 4.1 apresenta, de maneira resumida, alguns números sobre os dados teóricos e experimentais de XL-MS dos quatro sistemas estudados, enquanto a tabela 4.2 apresenta brevemente as distâncias consideradas para cada *cross-link*.

Tabela 4.1: Sumário dos dados teóricos e experimentais obtidos pada cada proteína modelada

Proteína Modelada	SalBIII	HSA-D1	HSA-D2	HSA-D3
Comprimento da Sequência Primária	134	201	189	193
Aminoácidos acessíveis ao solvente	132	198	189	193
Número total de pares de aminoácidos	8911	20100	17766	18528
Pares de resíduos reativos na estrutura primária	630	2080	2145	1380
Pares reativos / Pares totais	0,071	0,010	0,012	0,074
Número de restrições cristalográficas teóricas*	74	167	181	119
Restrições experimentais obtidas	156	163	195	98
Restrições experimentais consistentes com a estrutura cristalográfica	29	46	55	34
Restrições triviais	9	11	11	13
Restrições cristalográficas e não-triviais (VP)	20	35	44	21
Restrições por resíduo	0,149	0,174	0,233	0,109
TVP (Sensibilidade) versus total experimental	0,128	0,215	0,226	0,214
TVP (Sensibilidade) versus total teórico	0,270	0,210	0,243	0,176

* - Computada utilizando a estrutura cristalográfica e as distâncias da tabela 4.2

Tabela 4.2: Distâncias consideradas para cada tipo de *cross-link*

Tipo de *cross-link*	Distância utilizada
GLU(C_β)—GLU(C_β)	16,7 Å
GLU(C_β)—ASP(C_β)	15,4 Å
ASP(C_β)—ASP(C_β)	14,1 Å
ASP(C_β)—LYS(C_β)	9,6 Å
GLU(C_β)—LYS(C_β)	10,3 Å
GLU(C_β)—SER(C_β)	7,1 Å
ASP(C_β)—SER(C_β)	6,2 Å
MET(C_β)—LYS(C_β)	18,0 Å
LYS(C_β)—LYS(C_β)	21,8 Å
LYS(C_β)—SER(C_β)	18,0 Å
SER(C_β)—SER(C_β)	14,1 Å

Observa-se que existe certa heterogeneidade entre o número de *cross-links* possíveis, a quantidade de dados experimentais obtidos e o número de restrições compatíveis com a estrutura cristalográfica para cada uma da quatro proteínas. A partir da observação da tabela 4.1, é razoável supôr que cada domínio apresentará uma dificuldade diferente para a modelagem, sendo que espera-se, *a priori*, baseando-se na informação por resíduo e nas sensibilidades, que a modelagem do domínio HSA-D2 apresente os melhores resultados, logo em seguida as modelagens da SALBIII e HSA-D1 apresentem resultados igualmente intermediários (porque embora os dados da HSA-D1 sejam levemente melhores, a sequência da SALBIII é mais curta) e o domínio HSA-D3 seja o mais difícil de modelar com boa qualidade.

4.1.2 Quantas restrições recuperar?

Não existe, até o momento, uma estratégia bem definida e robusta para a seleção de dados para modelagem assitida por XL-MS que permitam auxiliar nessa escolha. No entanto, existem trabalhos que abordam a robustez da modelagem estrutural assistida por restrições de distância de forma geral. Estes estudos procuram estabelecer qual é o número mínimo de contatos entre resíduos que é necessário para a definição da topologia da estrutura. Os contatos são calculados a partir das estruturas cristalográficas, portanto são distâncias ideais do ponto de vista da modelagem, e são distâncias precisas. A qualidade e a tolerância da informação contida em dados de contato é diferente daquela prevista pelos dados de XL-MS, sendo que os dados de contato são muito mais precisos e restritivos.

Uma publicação do próprio David Baker (autor do Rosetta), datada de 2014, diz que, se a informação de estrutura secundária for perfeita, basta 1 restrição a cada 12 resíduos da sequência primária [86] para modelagem precisa e robusta da topologia proteica. Mais recentemente, em 2018, Mandalaparthy e colaboradores [87] resgatam essa informação e vão além, dizendo que o número de restrições necessárias para realizar uma modelagem é alguma proporção entre 5 e 10% do mapa de contatos nativo de uma proteína. Uma vez que, no presente trabalho, a informação de estrutura secundária não é perfeita, e que os dados de XL-MS são muito menos informativos, esses valores foram entendidos como limites inferiores para a quantidade de dados a serem recuperados.

De fato, nesse trabalho foi proposta a recuperação de um determinado número **M** de restrições, cuja relação com o tamanho **N** da sequência primária é $M = 0,3 \times N$, onde M é truncado para o inteiro mais próximo. De acordo com essa proposição, serão recuperadas 40, 60, 57 e 58 restrições, respectivamente, nos sistemas SALBIII, HSA-D1, HSA-D2 e HSA-D3.

4.1.3 Parâmetros de execução do Rosetta

Para todos os 4 sistemas estudados, as opções de execução do rosetta foram as seguintes:

- Protocolo *abinitio* com as opções `-fastrelax -increase_cycles 5 -rg_reweight 0.25`
- 5000 modelos gerados no caso da modelagem preliminar e 500 modelos a cada iteração
- Gravação dos modelos atomísticos no formato *silent* com as opções `-file -fullatom -silent folding_silent.out`
- Introdução do arquivo de personalização `xl` com as opções `-constraints -cst_file $xl -cst_weight 1 -cst_fa_file $xl -cst_fa_weight 1`

4.2 Modelagens da SALBIII

4.2.1 Recuperação frequentista

A recuperação frequentista, após dez iterações sucessivas do protocolo de modelagem, resultou num conjunto em que nenhum modelo atingiu TM-score acima de 0,5 em relação à estrutura

cristalográfica. Um máximo é atingido na quarta iteração, contudo não representa incremento expressivo na qualidade. A figura 4.1 apresenta as distribuições de qualidade para todos os modelos gerados. Nela, a região sombreada em azul, correspondente a TM-score acima de 0,5 em relação à estrutura cristalográfica, representa a zona em que os modelos são considerados bem-sucedidos em termos de topologia compatível com PDB-5CXO; as cores mais claras representam iterações iniciais e as cores mais escuras representam iterações finais.

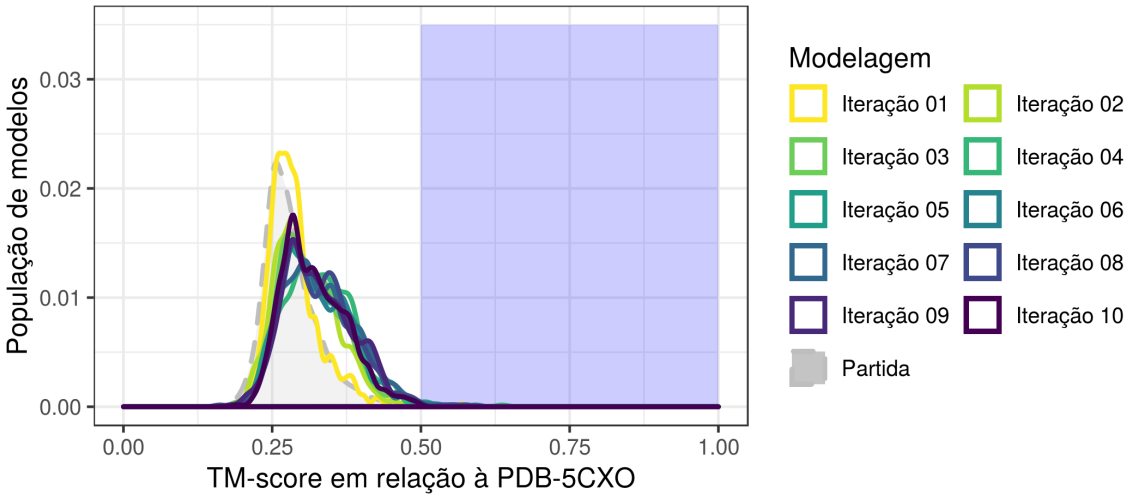

Figura 4.1: Distribuições de qualidade dos modelos para as 10 iterações do protocolo de modelagem aplicadas à proteína SALBIII com recuperação de restrições pelo critério frequentista.

Conforme já afirmado, esse critério foi inicialmente proposto - além da comparação com a literatura [43] - por um conjunto de razões que o tornam prático, já que não depende do cômputo de qualquer outro escore, podendo ser determinado simples e unicamente a partir de uma matriz binária que computa a obediência de cada modelo a cada restrição.

Contudo, a qualidade genérica das modelagens, que mal alcançam 1% dos modelos com qualidade aceitável, contrasta com a constatação de que mais de 50% das 40 restrições recuperadas estão factualmente presentes na estrutura cristalográfica. Chama-se atenção aqui novamente para o conceito de restrições triviais, incluído na seção 2.4.2, uma vez que confirma-se a recuperação constante de todas as 9 restrições cristalográficas e triviais do sistema SALBIII como as primeiras colocadas em todas as modelagens. De fato, a consistência do conjunto de 40 restrições ao longo das iterações foi surpreendentemente alta, levando à recuperação sistemática de um conjunto imutável de mais de 30 das 40 restrições.

4.2.2 Recuperação baseada no coeficiente bisserial empregando como variável contínua a similaridade à estrutura cristalográfica

A recuperação baseada no coeficiente bisserial e usando como referência a estrutura cristalográfica, após dez iterações sucessivas do protocolo de modelagem, resultou num conjunto de em que 23% dos modelos atingiram TM-score acima de 0,5 em relação à estrutura nativa. A figura 4.2 apresenta as distribuições de qualidade para todos os modelos gerados. Nela, a região sombreada em azul, correspondente a TM-score acima de 0,5 em relação à estrutura cristalográfica, representa a zona em que os modelos são considerados bem-sucedidos em termos de topologia compatível com PDB-5CXO; as cores mais claras representam iterações iniciais e as cores mais escuras representam iterações finais.

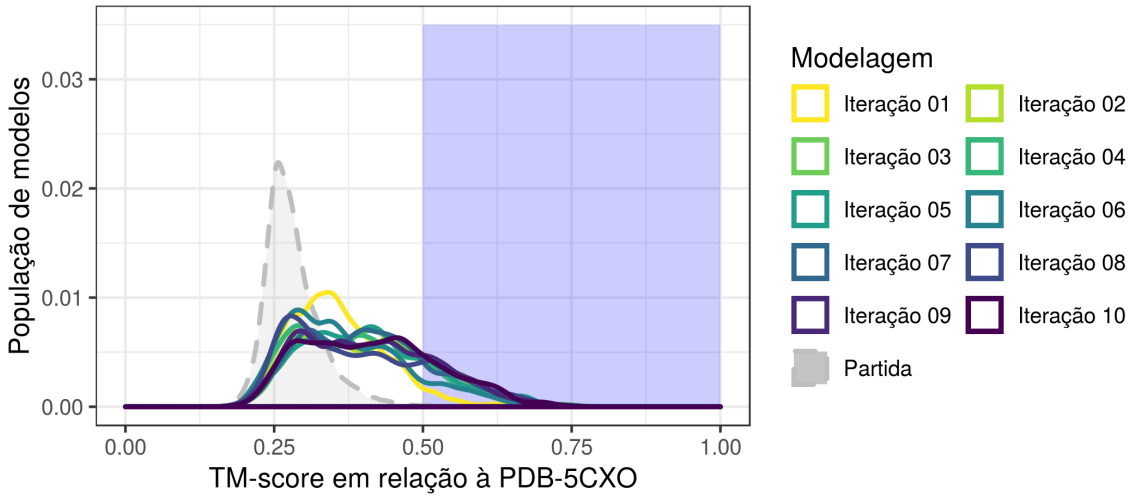

Figura 4.2: Distribuições de qualidade dos modelos para as 10 iterações do protocolo de modelagem aplicadas à proteína SALBIII com recuperação baseada no coeficiente bisserial empregando como variável contínua a similaridade à estrutura cristalográfica

Esse critério foi proposto como um caso ideal, ou *benchmark* do coeficiente de correlação ponto-bisserial. Se a qualidade de cada modelagem está sendo medida por meio da similaridade com a estrutura cristalográfica, então parece lógico que os melhores resultados sejam alcançados quando todos os modelos são artificialmente comparados a ela. Certamente essa não é uma reprodução de um caso real do problema de modelagem, mas permite aferir um limite superior ao desempenho do indicador de qualidade.

A substituição do critério frequentista permitiu um sensível incremento na qualidade dos

modelos gerados e no número de restrições cristalográficas não-triviais recuperadas a cada passo da iteração. Esse resultado endossa a hipótese levantada na seção 2.4.2, segundo a qual não basta recuperar boa proporção de restrições cristalográficas, mas é também importante que essas restrições sejam não-triviais.

Comparando, por exemplo, a quarta iteração de cada um dos resultados apresentados, enquanto no caso da recuperação frequentista há 22 restrições nativas recuperadas num conjunto que gera 1% de modelos de boa qualidade, o caso ideal do coeficiente bisserial proposto permitiu a recuperação de 20 restrições cristalográficas (que seria considerado um conjunto pior), produzindo, no entanto, 18,2% de modelos de boa qualidade (uma modelagem muito melhor). A aparente discrepância é resolvida ao comparar a proporção de restrições triviais e não-triviais que foram utilizadas em cada modelagem. Enquanto, na seleção baseada na frequência, 9 das 22 restrições são triviais, na nova seleção realizada foram recuperadas apenas 5 restrições triviais, de modo que as 15 restrições nativas não-triviais restantes performam muito superiormente às 13 de outrora. Ressalta-se que esse fenômeno é consequência não só do aumento da proporção de restrições cristalográficas, mas também porque essas restrições são mais diferenciadoras de modelos de melhor qualidade.

Esse resultado foi importante para confirmar o quanto é promissor o coeficiente de correlação ponto-bisserial combinado a uma escala que efetivamente contabiliza a qualidade de cada modelo.

4.2.3 Recuperação baseada no coeficiente bisserial empregando como variável contínua a similaridade ao melhor modelo

A recuperação baseada no coeficiente bisserial e usando como referência o melhor modelo gerado a cada passo, após dez iterações sucessivas do protocolo de modelagem, resultou num conjunto de em que 21% dos modelos atingiram TM-score acima de 0,5 em relação à estrutura nativa. Um máximo é atingido na sexta iteração, contudo procede uma queda de qualidade, que será melhor analisada na seção 4.2.5. A figura 4.3 apresenta as distribuições de qualidade para todos os modelos gerados.

Esse critério é uma flexibilização do caso ideal proposto na aplicação anterior *benchmark* do coeficiente de correlação ponto-bisserial. Na eventualidade de não existir uma

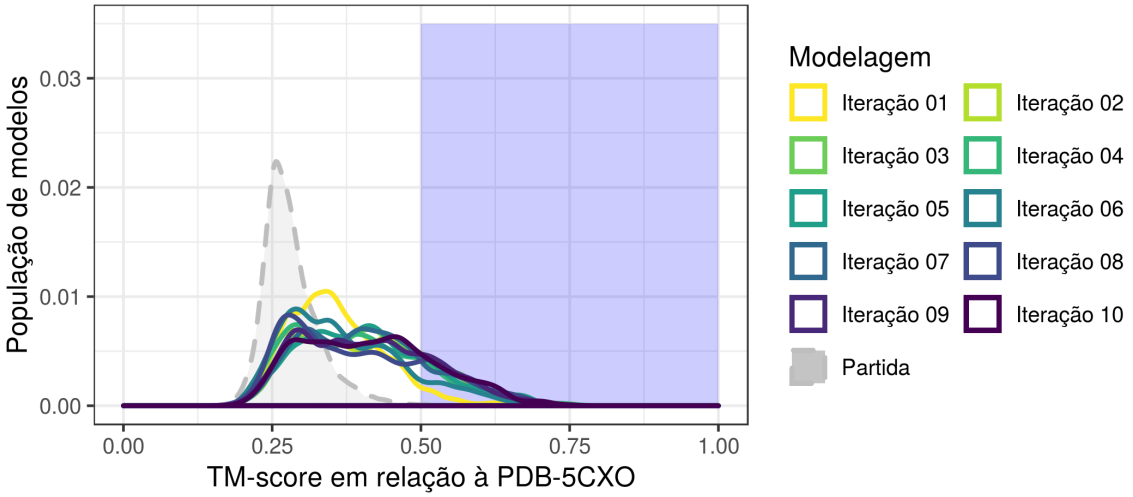

Figura 4.3: Distribuições de qualidade dos modelos para as 10 iterações do protocolo de modelagem aplicadas à proteína SALB3 com recuperação de restrições baseada no coeficiente bisserial empregando como variável contínua a similaridade ao melhor modelo.

estrutura cristalográfica para guiar a modelagem, propõe-se escolher sempre como referência o melhor modelo obtido no passo anterior. A hipótese era que, dada uma ferramenta ideal de seleção de modelos, a estrutura de referência poderia partir da própria modelagem, e, se fosse esse um modelo de alta qualidade – ou seja, de alta similaridade em relação à estrutura cristalográfica -, poderia substituí-la no protocolo sem perda de correspondência. Evidentemente isso faz o protocolo depender de ser sempre gerado pelo menos um modelo de topologia adequada a cada passo (mas principalmente na modelagem inicial) para permitir o progresso genérico.

Da mesma forma como foi feito para a subseção 4.2.2, pode ser novamente feita uma comparação entre a quarta iteração de cada um dos resultados apresentados: novamente o critério frequentista recupera 22 restrições cristalográficas, sendo 13 não-triviais, e alcança 1% de modelos de boa qualidade; já o critério aqui empregado - recuperação baseada no melhor modelo - recupera 20 restrições cristalográficas, sendo 15 não-triviais, alcançando 17% de modelos de boa qualidade.

4.2.4 Recuperação baseada no coeficiente bisserial empregando como variável contínua a similaridade ao modelo de maior ProQ3D-tmscore

A recuperação baseada no coeficiente bisserial e usando como referência um modelo selecionado pelo ProQ3D, um classificador independente de modelos, após dez iterações sucessivas do protocolo de modelagem, resultou num conjunto de em que 10% dos modelos atingiram TM-score acima de 0,5 em relação à estrutura cristalográfica. Um máximo é atingido na oitava iteração, contudo procede uma queda de qualidade, que será melhor analisada na seção 4.2.5. A figura 4.4 apresenta as distribuições de qualidade para todos os modelos gerados.

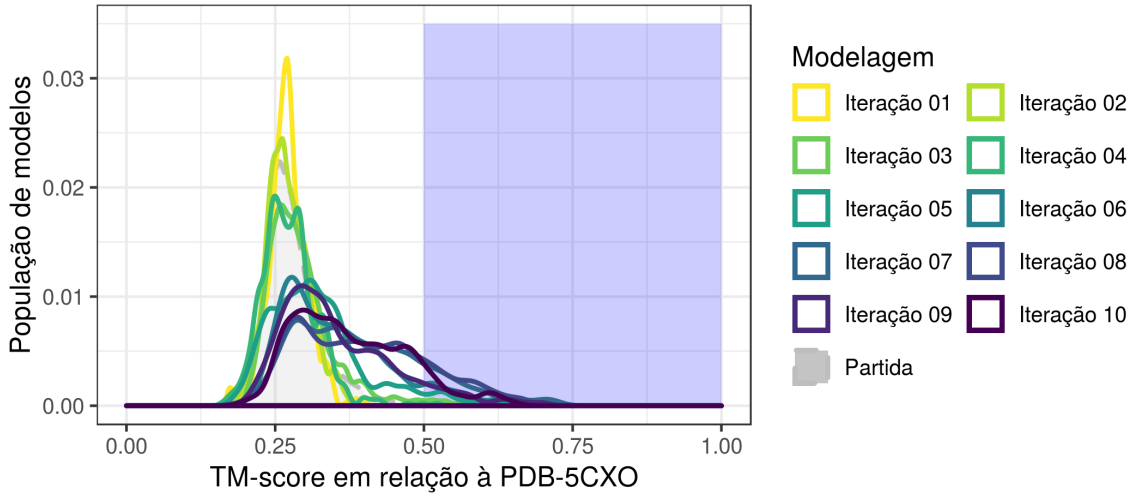

Figura 4.4: Distribuições de qualidade dos modelos para as 10 iterações do protocolo de modelagem aplicadas à proteína SALB3 com recuperação de restrições pelo critério bisserial empregando como variável contínua a similaridade ao modelo de maior ProQ3D-tmscore

O melhor modelo selecionado pelo ProQ3D mostrou-se consistentemente capaz de alavancar a qualidade das modelagens, e, conforme esperado, tem um desempenho levemente inferior à seleção artificial do melhor modelo (uma vez que nem sempre o modelo de maior ProQ3D-tmscore - calculado por meio de uma regressão - é, de fato, o mais similar à estrutura cristalográfica). A décima e última iteração realizada alcançou 10% de modelos com TM-score acima de 0.5 em relação à estrutura cristalográfica, gerando modelos com similaridade de até 0,746 com a nativa. Para a última iteração, foram recuperadas 20 restrições cristalográficas, das quais 15 eram não-triviais, ressaltando mais uma vez o mérito do coeficiente ponto-bisserial em

filtrar as restrições triviais.

Nesse protocolo, portanto, que representa um caso real (em que não se pode contar com a existência de uma estrutura cristalográfica para selecionar o modelo mais similar) que combinou as características da correlação bisserial à boa seleção de modelos realizada pelo ProQ3D, foi alcançado um incremento de mais de 30 vezes no número de modelos de boa qualidade, em relação à modelagem de partida que continha apenas 0,32% de bons modelos. paralelamente à recuperação de 40% mais restrições não-triviais no final do protocolo em relação à recuperação frequentista.

4.2.5 Resumo e Discussões das modelagens para a SALBIII

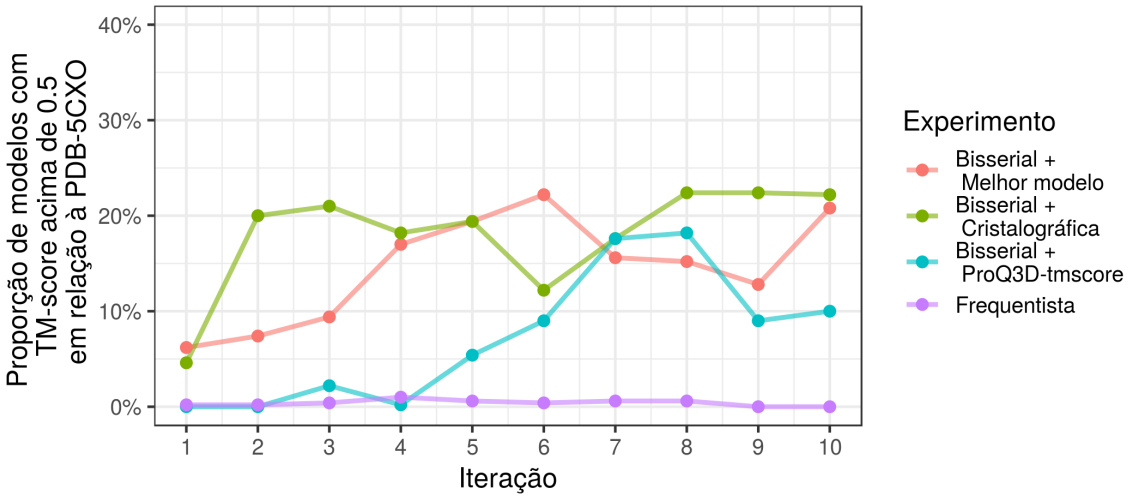

Figura 4.5: Acompanhamento da qualidade das modelagens ao longo das dez iterações de modelagem da proteína SALB3 realizadas, para cada um dos critérios de seleção das restrições.

A Figura 4.5 apresenta um gráfico de linhas que permite acompanhar o progresso das modelagens para cada critério de recuperação. A proteína SALBIII apresentou, em linhas gerais, os resultados esperados da metodologia, ou seja, a recuperação frequentista com baixo desempenho, em detrimento da recuperação com coeficiente bisserial baseada tanto na estrutura cristalográfica quanto no melhor modelo de cada passo (ou seja, o mais similar à cristalográfica), que apresentaram desempenho genericamente crescente ao longo das iterações, culminando em mais de 20% de modelos bem-sucedidos. Já a modelagem que simula o caso real onde a estrutura cristalográfica não pode ser utilizada como referência para seleção dos modelos atinge um desempenho intermediário, mas ainda assim muito superior à recuperação frequentista, gerando

10% de bons modelos ao final de 10 iterações.

No caso das iterações onde se observa uma pequena queda de qualidade em relação ao passo anterior, algumas hipóteses foram traçadas do porque isso ocorre. A primeira delas foi que havia instabilidade na qualidade do melhor modelo selecionado a cada iteração, e que essa instabilidade se refletiria na qualidade geral da modelagem.

Essa instabilidade é esperada, e advém do fato de que muitas vezes a produção um único modelo de boa qualidade depende muito mais dos aspectos probabilísticos da modelagem como, por exemplo, a seleção dos fragmentos adequados e o progresso da simulação estocástica, do que do conjunto de restrições fornecido à modelagem. É por isso, por exemplo, que se modela um grande conjunto de modelos a cada iteração.

Um exemplo claro disso é a comparação das iterações 06, 07 e 08 do critério baseado no coeficiente bisserial empregando como variável contínua a similaridade ao melhor modelo: a iteração 06 produz $22{,}2\%$ de modelos bem-sucedidos, onde o melhor modelo tem TM-score em relação à estrutura cristalográfica de $0{,}6913$; já a iteração seguinte - 07 - produz $15{,}6\%$ de modelos bem-sucedidos (uma queda de desempenho, enquanto o melhor modelo gerado é melhor que o da modelagem anterior, apresentando TM-score de $0{,}7075$; nesse caso, seria esperado um incremento de qualidade no passo posterior, mas a iteração 08 produz como melhor modelo um de TM-score $0{,}6877$ e uma porcentagem de modelos bem-sucedidos de $15{,}2\%$.

Isso mostrou que não necessariamente existe uma grande correlação entre a qualidade da modelagem e a qualidade do melhor modelo. Essa constatação fica ainda mais evidente quando se percebe que existem quedas de desempenho também nas modelagens que utilizam como referência a restrição cristalográfica. Se os únicos critérios relevantes fossem a qualidade do modelo de referência e a qualidade da modelagem anterior, esse protocolo deveria apresentar um crescimento monotônico de qualidade até atingir um platô, mas não é isso que acontece. Isso levou a uma das conclusões parciais desse projeto, que é que a qualidade do modelo selecionado como referência é um dos principais fatores do sucesso da modelagem, mas não é o único.

Logo, outra hipótese foi traçada, que foi a que o tamanho do conjunto de restrições recuperado, em especial o viés introduzido pelos falsos-positivos recuperados, tem tanta importância quanto a qualidade do modelo de referência. De fato, é importante não somente a taxa de falsos

positivos observada a cada iteração, mas também a cooperação entre esses falsos positivos: se todos eles apontarem para conformações diferentes, as restrições cristalográficas irão ponderar, mas do contrário, havendo cooperação entre eles no sentido de amostrar outra conformação, eles podem levar à diminuição do viés em relação à estrutura cristalográfica.

Para exemplificar esse fenômeno, ainda no ínterim dos resultados obtidos para o critério BISCORE-BEST, construiu-se a projeção do espaço conformacional para as iterações 05 a 10, que pode ser visualizada na figura 4.6. Nessa figura, os pontos estão coloridos por densidade de modelos na projeção, e um círculo em cinza foi adicionado para mostrar a vizinhança da estrutura cristalográfica.

De fato, a partir das iterações 05 e 06, começam a aparecer claramente duas conformações distintas: uma delas, mais similar á estrutura cristalográfica, assinalada em verde; outra, mais distante, assinalada em vermelho.

Hipotetizou-se que, nesses casos, alguns falsos-positivos estariam colaborando entre si para amostragem de duas conformações, sendo uma delas de topologia mais adequada do que a outra. De fato, a distribuição dos modelos entre essas duas conformaçõe é a principal responsável para que, a partir da iteração 07, observe-se uma queda na qualidade geral da modelagem, que pode ser conferida na figura 4.6. Esse fenômeno de distribuição entre os dois estados só é revertido para a conformação principal na décima iteração.

Para tentar entender melhor o que estava acontecendo, produziu-se a Figura 4.7, representando o alinhamento tridimensional entre a estrutura cristalográfica e duas estruturas representativas de cada uma das conformações observadas. Nessa figura, observa-se em laranja a estrutura cristalográfica, em verde a estrutura representativa da conformação mais similar e, em vermelho, a estrutura mais representativa da conformação menos similar, conforme as cores assinaladas na Figura 4.6.

Existem principalmente duas zonas de discordância entre as estruturas. A primeira zona é o posicionamento relativo das α-hélices α_1, α_2 e α_3. Na conformação em vermelho, parece ocorrer um distanciamento da hélice α_3 em relação ao restante da proteína, penalizando o TM-score da estrutura vermelha. A segunda zona é o *loop* existente entre as seções de folha-β β_5 e β_6. Ambas as conformações amostradas na modelagem parecem divergir da estrutura cristalográfica,

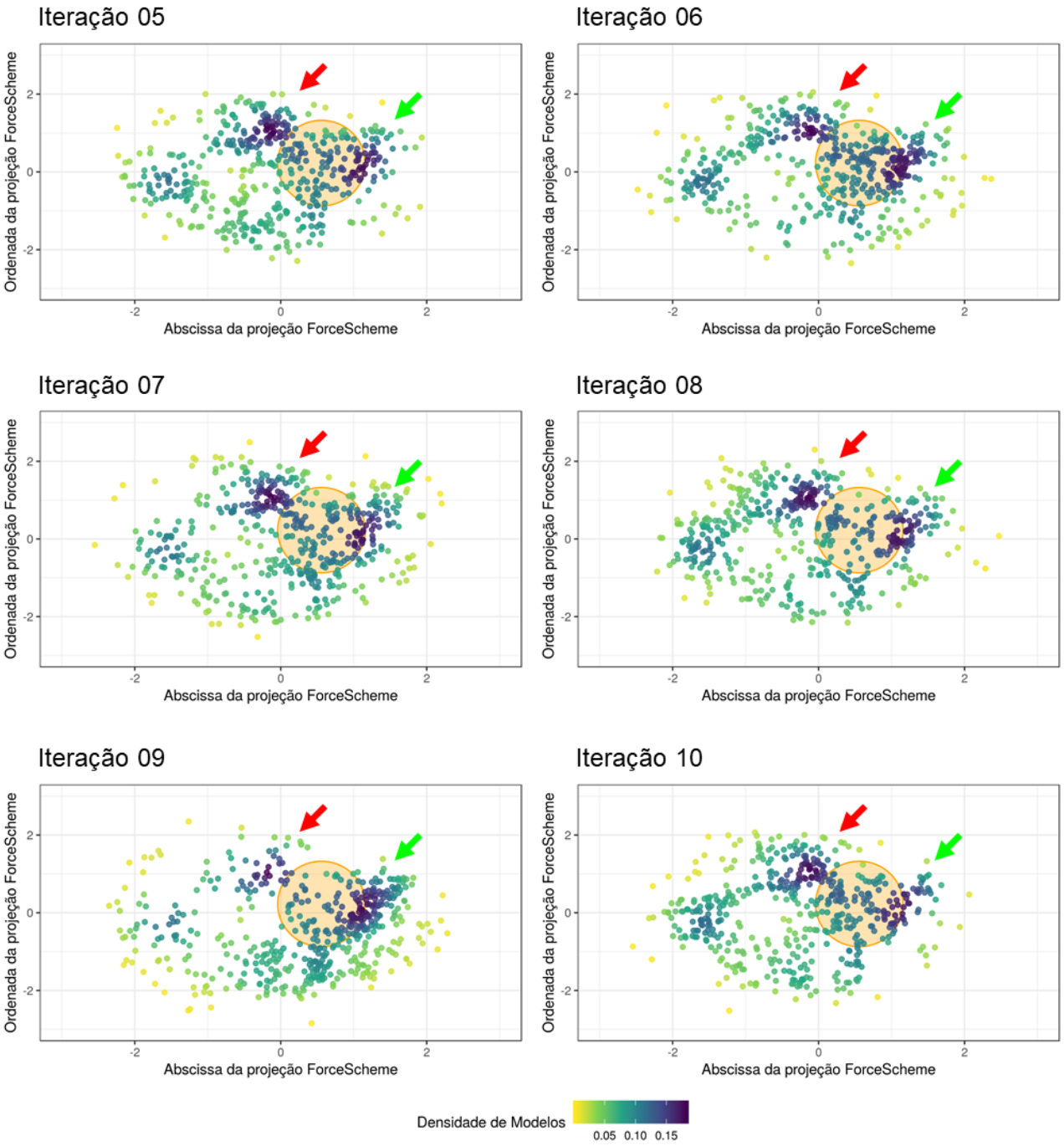

Figura 4.6: Evolução das projeções do espaço conformacional, computado mediante algoritmo ForceScheme, para as iterações 05 a 10 da modelagem da proteína SALBIII com recuperação de restrições baseada no coeficiente bisserial empregando como variável contínua a similaridade ao melhor modelo.

mas na conformação em vermelho parece haver uma perda mais sensível na informação de estrutura secundária. Esses fenômenos infelizmente não podem relacionados à seleção de restrições, visto que,

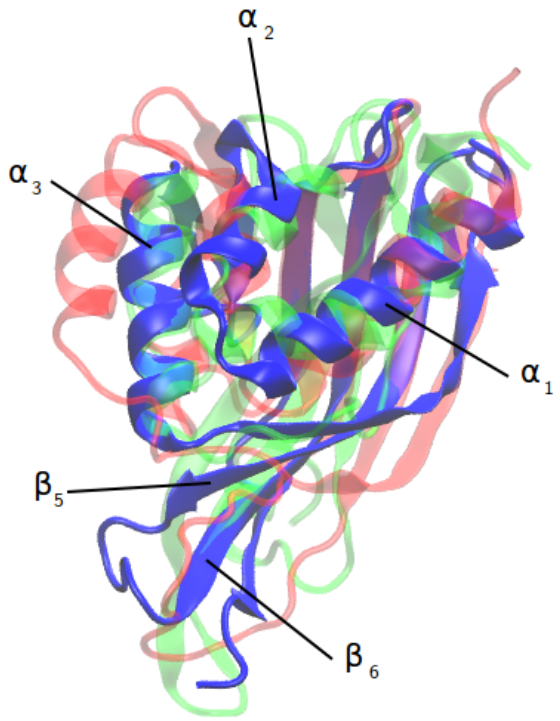

Figura 4.7: Alinhamento entre a estrutura cristalográfica (PDB-5CXO, em azul) e as estruturas representativas de ambas as conformações amostradas conforme a Figura 4.6.

quando se observa novamente a Figura 2.2, nota-se que não foi possível aferir experimentalmente nenhuma restrição capaz de cobrir essas porções da proteína.

Alguns tipos de instabilidade na qualidade da modelagem podem ser atenuados pela diminuição do número de restrições recuperadas a cada passo iterativo, uma vez que a taxa de falsos positivos (e, portanto, a probabilidade desse fenômeno ocorrer) aumenta genericamente com o aumento do número de restrições recuperadas. Esse fenômeno será determinante para alguns resultados observados no caso da HSA-D3, mais adiante. No entanto, no caso da modelagem da SALBIII aqui discutida, fica clara principalmente a colaboração negativa da qualidade dos dados experimentais.

Isso não invalida a principal conclusão parcial dessa etapa do trabalho: para a proteína SALBIII, a seleção aplicando coeficiente bisserial produziu, em todos os três casos em que foi aplicada e num protocolo iterativo, uma modelagem de qualidade superior àquela que utiliza restrições recuperadas apenas pela frequência. A qualidade de cada modelagem é função principalmente de três fatores, a saber: a similaridade do modelo selecionado como referência no

coeficiente bisserial em relação à estrutura objetivada na modelagem; a qualidade dos dados experimentais e a informação estrutural que eles provêm; o tamanho do conjunto de restrições recuperado e a capacidade das restrições recuperadas de - sendo elas consistentes com a estrutura objetivada ou não - de diferenciarem entre modelos de maior ou melhor qualidade.

4.3 Resultados obtidos para os três domínios da HSA

4.3.1 Recuperação frequentista

A recuperação frequentista, após dez iterações sucessivas do protocolo de modelagem para cada domínio, produziu resultados similares para cada proteína. No caso da HSA-D1, em nenhuma iteração é produzida quantidade sensível de modelos com TM-score acima de 0,5 em relação à PDB-1AO6-D1. De fato, o máximo, alcançado na iteração 02, corresponde a apenas 1% de modelos bem-sucedidos. Para a HSA-D2 e HSA-D3, a situação é ainda pior, onde nenhuma iteração foi capaz de produzir sequer um modelo com TM-score acima de 0,5 em relação à PDB-1AO6-D2 e PDB-1AO6-D3, respectivamente.

De fato, a análise da Figura 4.8 mostra que, em todos os três casos, mesmo após 10 iterações, além de não ser observado incremento expressivo nas qualidades das modelagens, o formato da distribuição de qualidade das modelagens permaneceu essencialmente igual. Isso ocorre principalmente porque, de forma similar ao que aconteceu com a SALBIII, o conjunto de restrições recuperadas, que contém sempre uma combinação de muitas restrições triviais e outras restrições que não são particularmente discriminantes entre os modelos gerados (uma vez que nenhum tipo de medida de diferenciação é aqui aplicada), é incapaz de enviesar a modelagem em direção às estruturas cristalográficas de cada domínio.

4.3.2 Recuperação baseada no coeficiente bisserial empregando como variável contínua a similaridade à estrutura cristalográfica

Também foi realizado, para a HSA, o mesmo teste que foi feito na SALBIII, que consiste em aplicar o caso ideal do coeficiente bisserial, onde os escores de cada modelo são a própria

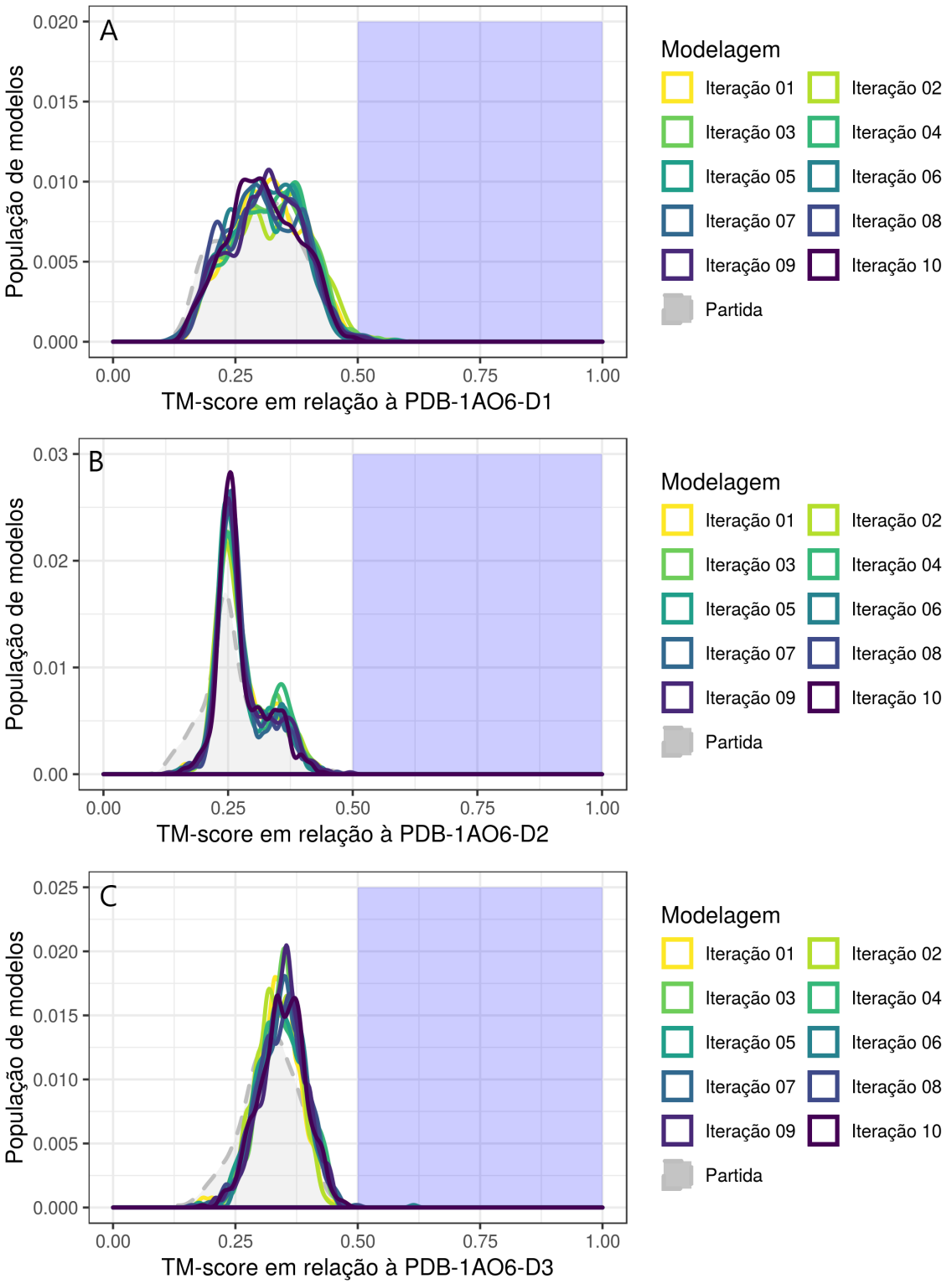

Figura 4.8: Distribuições de qualidade dos modelos para as 10 iterações do protocolo de modelagem aplicadas às proteínas HSA-D1 (A), HSA-D2 (B) e HSA-D3 (C) com recuperação de restrições pelo critério frequentista.

similaridade à estrutura cristalográfica. As distribuições resultantes estão na Figura 4.9.

Nesse caso, os resultados foram ligeiramente diferentes daqueles observados para a SALBIII. Para o domínio HSA-D1, observou-se muita instabilidade, com a maioria das iterações fornecendo menos de 5% de bons modelos, chegando em alguns casos a apenas 1%, mas se recuperando nas iterações 9 e 10 até alcançar um máximo de 10% de modelos bem-sucedidos.

No domínio HSA-D2, o comportamento foi aproximadamente ideal, apresentando majoritariamente um crescimento monotônico nas 8 primeiras iterações, sofrendo uma leve queda nas duas últimas, mas mesmo assim atingindo a surpreendente proporção 36,6% de modelos com TM-score acima de 0.5 em relação à estrutura cristalográfica. Se for comparada com a modelagem de partida, onde apenas 0,7% dos modelos alcançam essa marca, o incremento de qualidade foi superior a 50 vezes.

Entretanto, para o domínio HSA-D3, mesmo esse caso considerado ideal não foi capaz de produzir nenhum acréscimo considerável na qualidade das modelagens. Conforme já discutido e em observância com o que foi proposto na subseção 4.1.1, mesmo com uma seleção perfeita de modelo de referência, a baixa qualidade dos dados e a grande proporção de falsos positivos nos conjuntos recuperados impede o progresso da modelagem.

4.3.3 Recuperação baseada no coeficiente bisserial empregando como variável contínua a similaridade ao melhor modelo

Conforme pode-se observar pela análise da Figura 4.10, os resultados obtidos para esse critério foram, novamente e em linhas gerais, similares àqueles obtidos para o critério que emprega coeficiente bisserial e a similaridade à estrutura cristalográfica, ou seja: para os sistemas HSA-D1 e HSA-D2, essas modelagens superam muito o desempenho da modegalem com recuperação frequentista de restrições, atingindo respectivamente 13,2% (com um máximo de 16,2 na sétima iteração) e 21,4% (com um máximo de 34,2% na nona iteração) de modelos com TM-score acima de 0,5 em relação à estrutura cristalográfica. No caso do domínio HSA-D3, observou-se novamente que as modelagens não foram capazes de produzir nenhuma proporção significativa de modelos de boa qualidade.

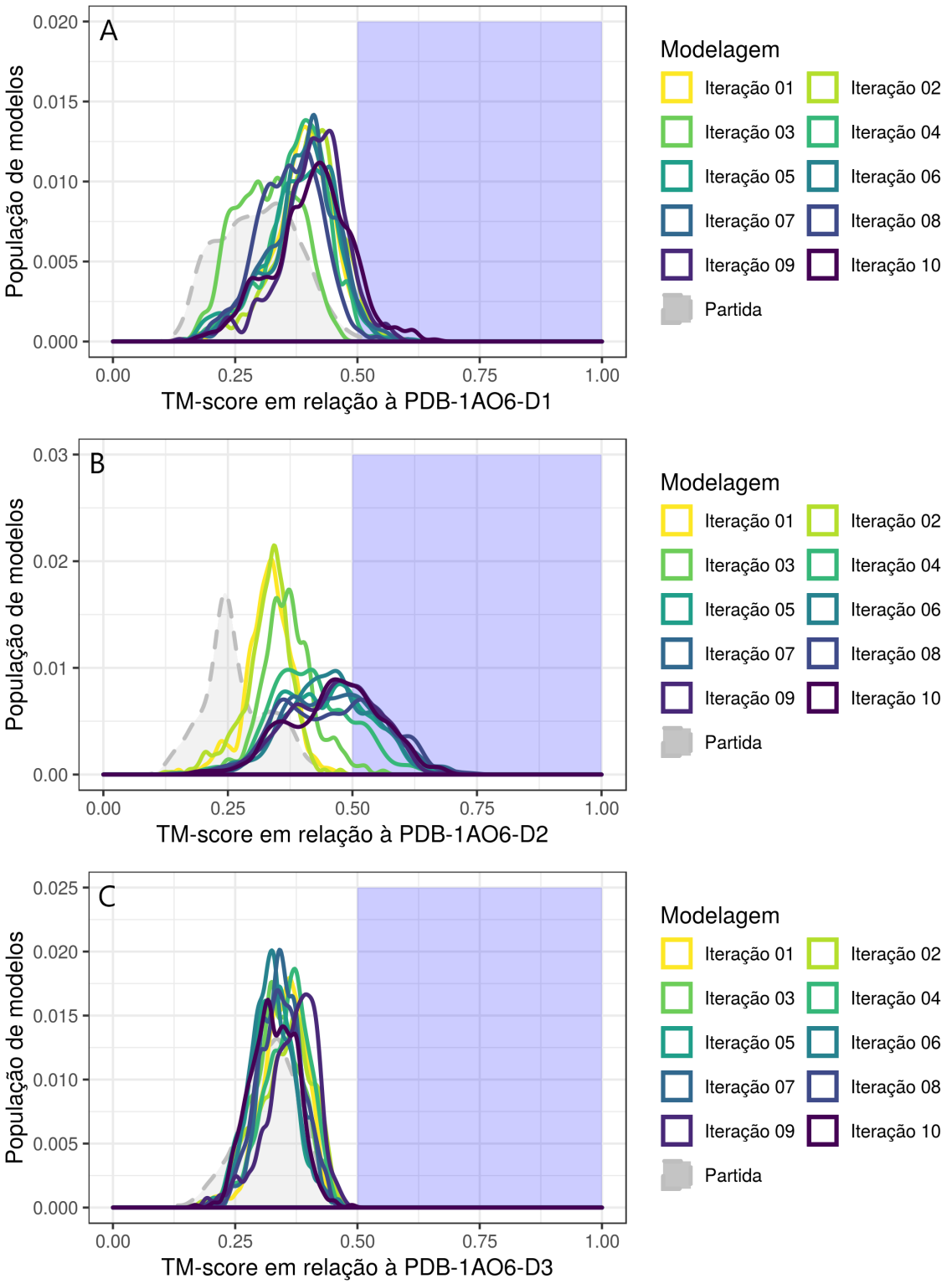

Figura 4.9: Distribuições de qualidade dos modelos para as 10 iterações do protocolo de modelagem aplicadas às proteínas HSA-D1 (A), HSA-D2 (B) e HSA-D3 (C) com recuperação de restrições baseada no coeficiente bisserial empregando como variável contínua a similaridade à estrutura cristalográfica.

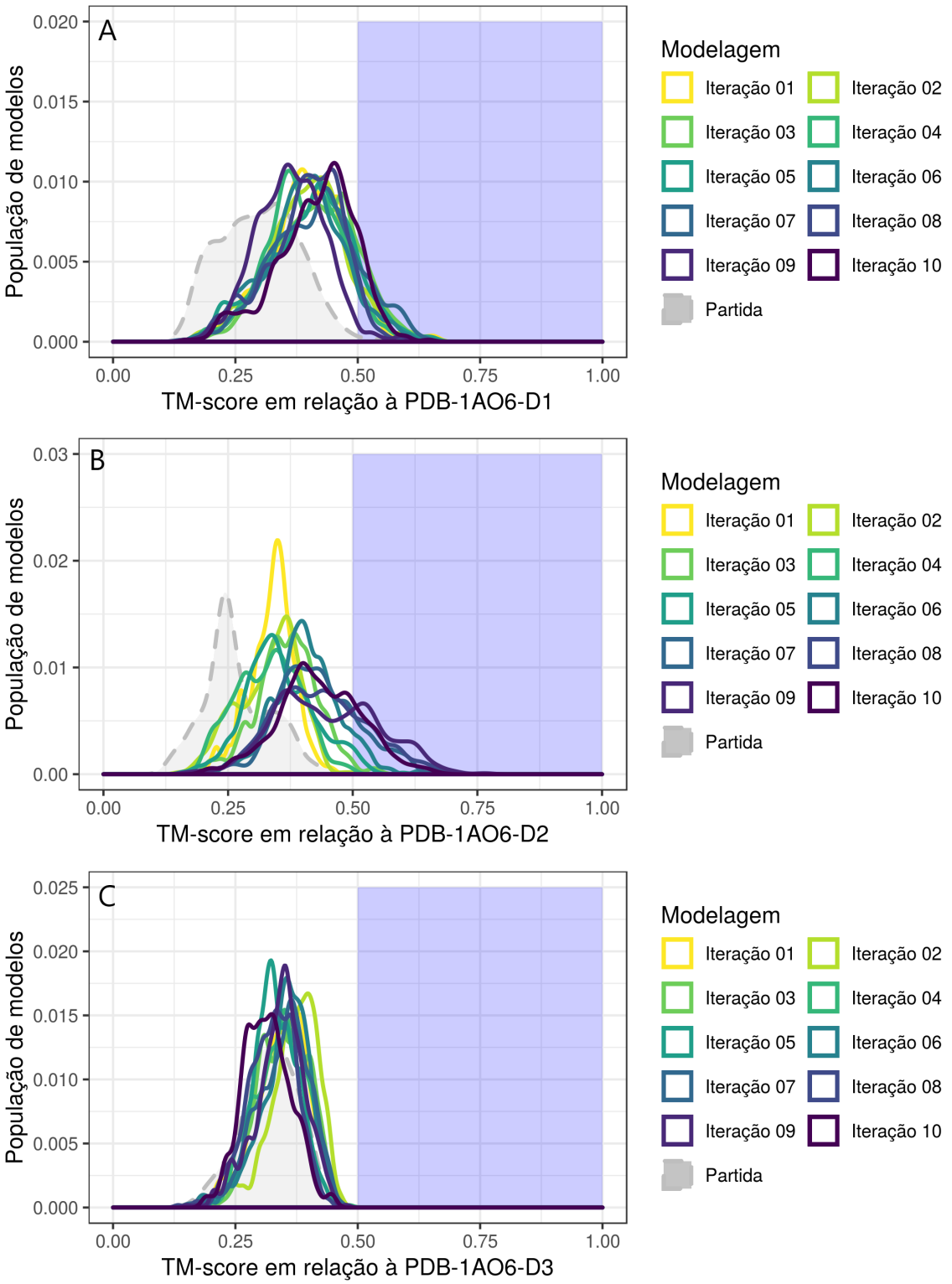

Figura 4.10: Distribuições de qualidade dos modelos para as 10 iterações do protocolo de modelagem aplicadas às proteínas HSA-D1 (A), HSA-D2 (B) e HSA-D3 (C) com recuperação de restrições baseada no coeficiente bisserial empregando como variável contínua a similaridade ao melhor modelo.

4.3.4 Recuperação baseada no coeficiente bisserial empregando como variável contínua a similaridade ao modelo de maior ProQ3D-tmscore

No caso da recuperação de restrições empregando o coeficiente bisserial e um modelo selecionado pelo ProQ3D-tmscore, os resultados foram muito semelhantes para todos os três domínios, em que praticamente nenhuma modelagem conseguiu gerar sequer mais de 1% de modelos de boa qualidade, com exceção da sétima iteração para a HSA-D1, que alcança um máximo de 2.5%. As distribuições obtidas para cada iteração realizada e para cada critério testado podem ser observadas na Figura 4.11.

Nesse caso, um dos principais motivos para o fracasso é o próprio desempenho do classificador ProQ3D. Para o primeiro passo da modelagem de cada um dos domínios, o TM-score do modelo selecionado na modelagem de partida em relação à estrutura cristalográfica é, respectivamente para os três domínios, 0,3679, 0,2818 e 0,3397. Esses modelos são tão dissimilares à estrutura cristalográfia que selecionar restrições que enviesem o espaço conformacional à vizinhança desses modelos é, na verdade, contraproducente do ponto de vista de produzir modelos similares à estrutura cristalográfica. A efeito de comparação, no caso da proteína SALBIII, onde as modelagens com o ProQ3D de fato avançam, o modelo selecionado como referência na modelagem de partida tem um TM-score de 0,5900 em relação à estrutura cristalográfica.

4.3.5 Resumo e discussões para HSA

Cabe aqui também, e com mais importância, uma discussão em relação ao tamanho do conjunto de restrições recuperado a cada passo de iteração. No caso da SALBIII, proteína de 134 aminoácidos, onde estavam sendo recuperadas 40 restrições, dentre as quais 20 a 25 eram cristalográficas, sendo 15 a 20 não-triviais, a taxa de falsos positivos recuperados nunca ultrapassava 0.5, à exceção das iterações iniciais.

Um panorama diferente é observado com os domínios da HSA. Conforme pode ser percebido pela leitura da tabela 4.1, de todas as restrições experimentais, apenas 46 no caso da HSA-

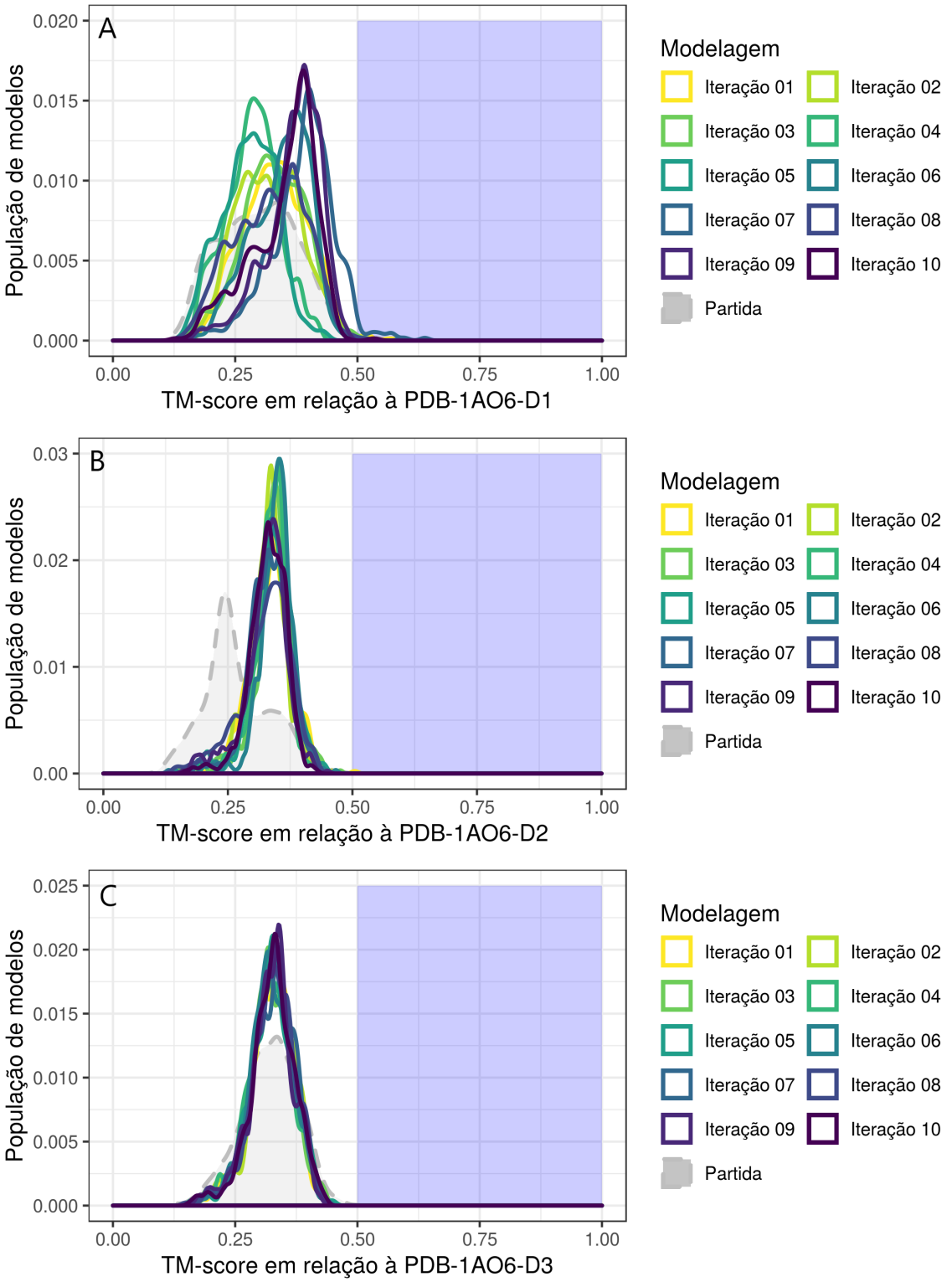

Figura 4.11: Distribuições de qualidade dos modelos para as 10 iterações do protocolo de modelagem aplicadas à proteína HSA-D3 com recuperação de restrições pelo critério BISCORE-PROQ3D.

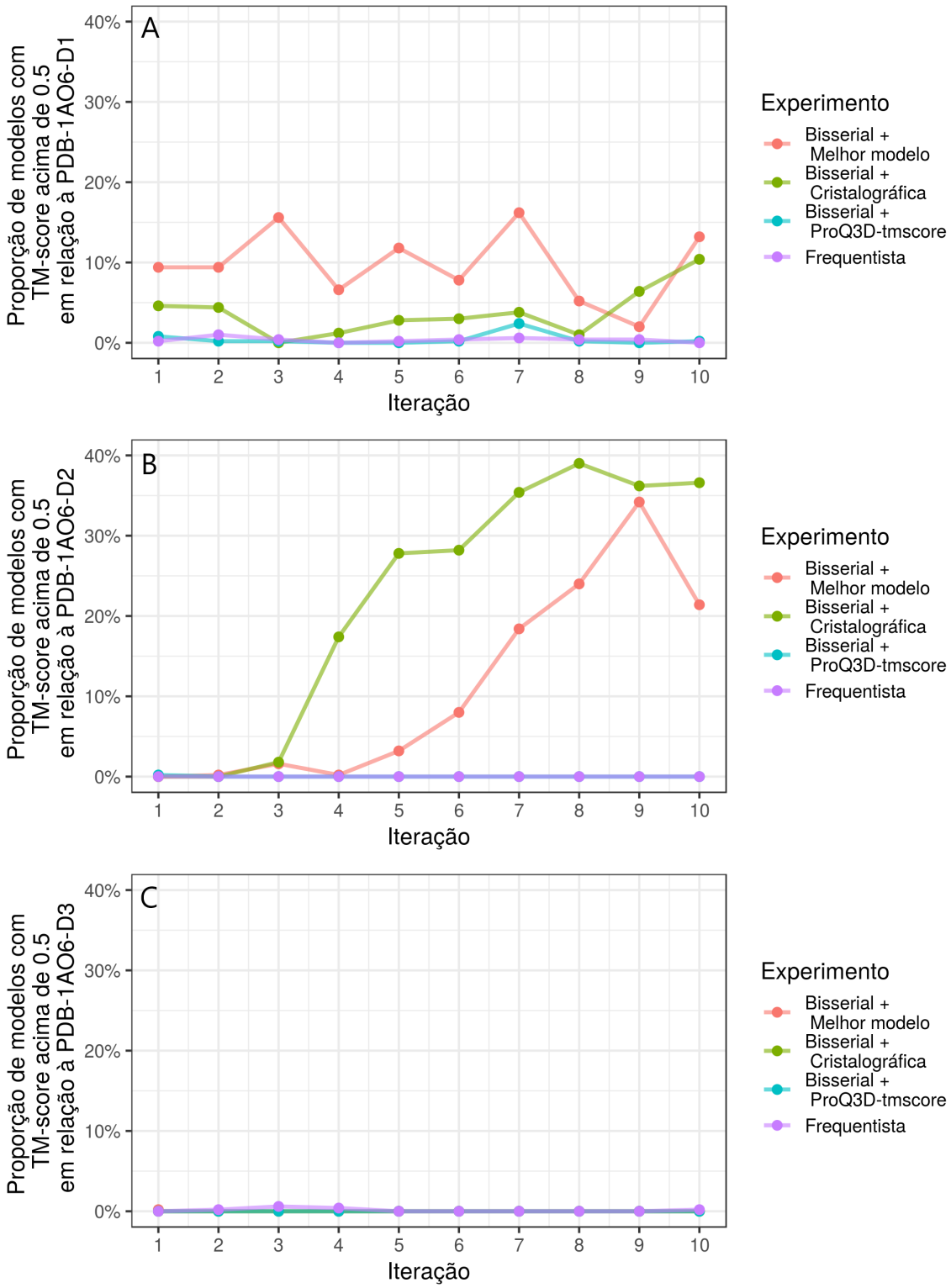

Figura 4.12: Acompanhamento da qualidade das modelagens ao longo das dez iterações de modelagem das proteínas HSA-D1 (A), HSA-D2 (B) e HSA-D3 (C) realizadas para cada um dos critérios de seleção das restrições.

D1, 55 no caso da HSA-D2 e 34 no caso da HSA-D3, são válidas perante a estrutura cristalográfica. Isso quer dizer, por exemplo, que, no caso da HSA-D3, onde são sempre recuperadas 58 restrições, no melhor cenário possível (em que são recuperadas todas as 34 restrições cristalográficas, sendo 21 não-triviais) a taxa de falsos positivos seria, no mínimo, 0,41, isso sem desconsiderar as restrições triviais. A consequência disso, conforme pode ser observado na análise do progresso das modelagens para a HSA na figura 4.12C, é que, no caso da HSA-D3, não foi observado nenhum progresso significativo no sentido de ampliar a taxa de modelos com topologia adequada nas iterações das modelagens.

Reexecução dos experimentos para HSA-D3

O que se observou ao longo de todas a iterações, para todos esses sistemas, foi valores em média muito mais altos para a taxa de falsos positivos nos conjuntos de restrições recuperados. Para ilustrar a importância novamente endossada desse fator, um experimento simples foi feito em que se modelou novamente a proteína HSA-D3, com todos os critérios já explorados, mas com recuperação de apenas 34 restrições em contraste com as 58 recuperadas outrora. Esse número foi sugerido por ser o exato número de restrições experimentais consistentes com a estrutura cristalográfica, conforme a Tabela 4.1. As distribuições de qualidade obtidas para esse experimento estão nas Figuras 4.13 e 4.14, e o progresso geral de modelagem pode ser encontrado na Figura 4.15

Pode-se observar que, quando o conjunto de restrições é diminuído em praticamente 40%, as modelagens *benchmark* utilizando como referência a estrutura cristalográfica e o melhor modelo conseguem, enfim, produzir conjuntos modelados onde observa-se uma proporção significativa de modelos de boa qualidade. Tomando por exemplo dois casos de modelagem, coeficiente bisserial empregando a similaridade à estrutura cristalográfica e coeficiente bisserial empregando a similaridade ao melhor modelo, as taxas de falsos positivos que oscilavam genericamente entre 50% e 60% na execução original agora foram diminuídas para um patamar entre 35% e 50%. As figuras de mérito comparativas para essas duas modelagens, na execução original e na reexecução do experimento, estão nas Tabelas 4.3 e 4.4. Portanto, adjunta-se às conclusões parciais obtidas na SALBIII um último fator determinante no progresso da modelagem como um todo, que é o tamanho do conjunto de restrições recuperado, que depende não só da sequência primária de proteínas mas também da quantidade e qualidade dos dados experimentais

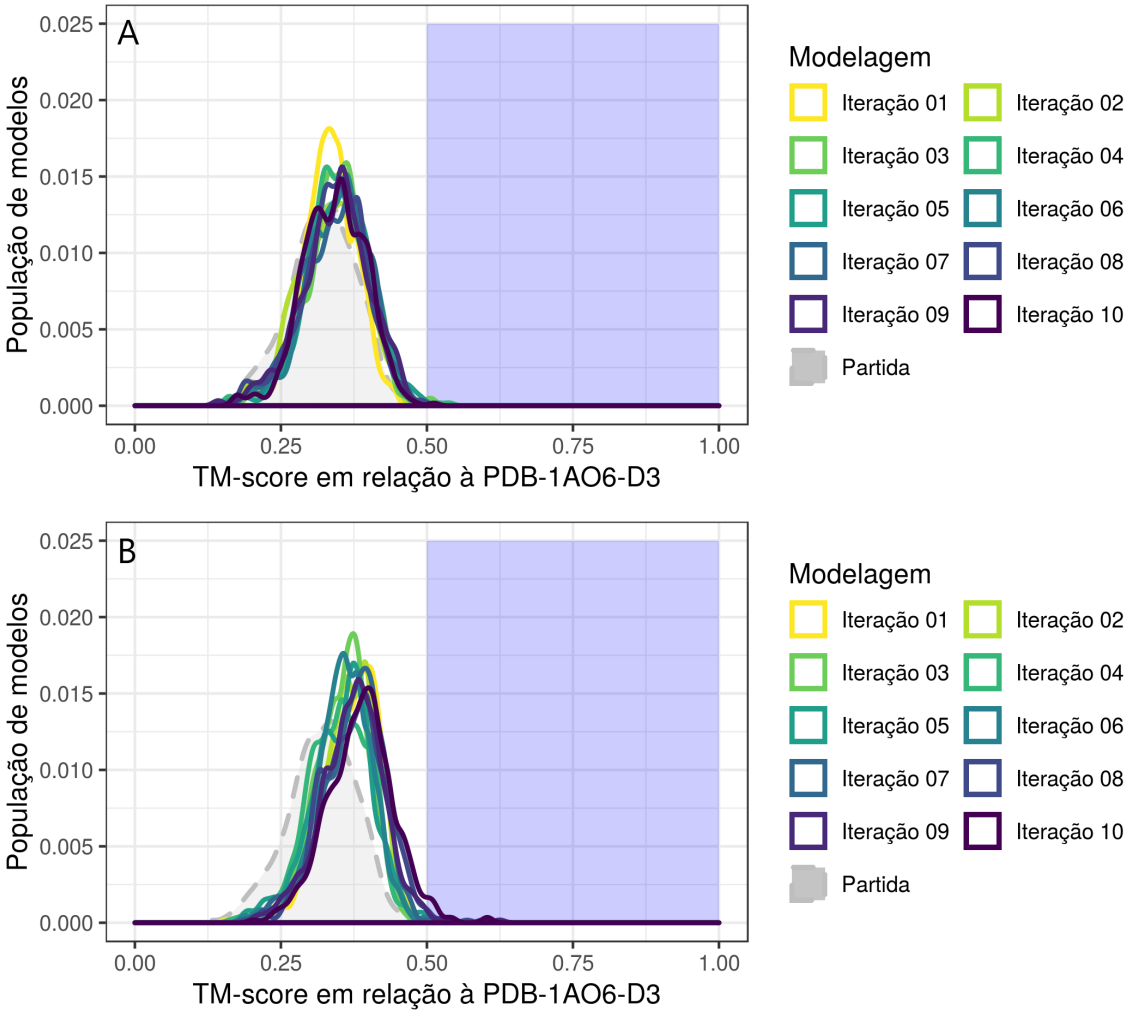

Figura 4.13: Progresso da qualidade das modelagens, na reexecução do experimento no sistema HSA-D3, com recuperação de 34 restrições e ao longo de dez iterações, para os quatro critérios já experimentados: frequentista (A) e coeficiente bisserial empregando a similaridade à estrutura cristalográfica (B)

obtidos.

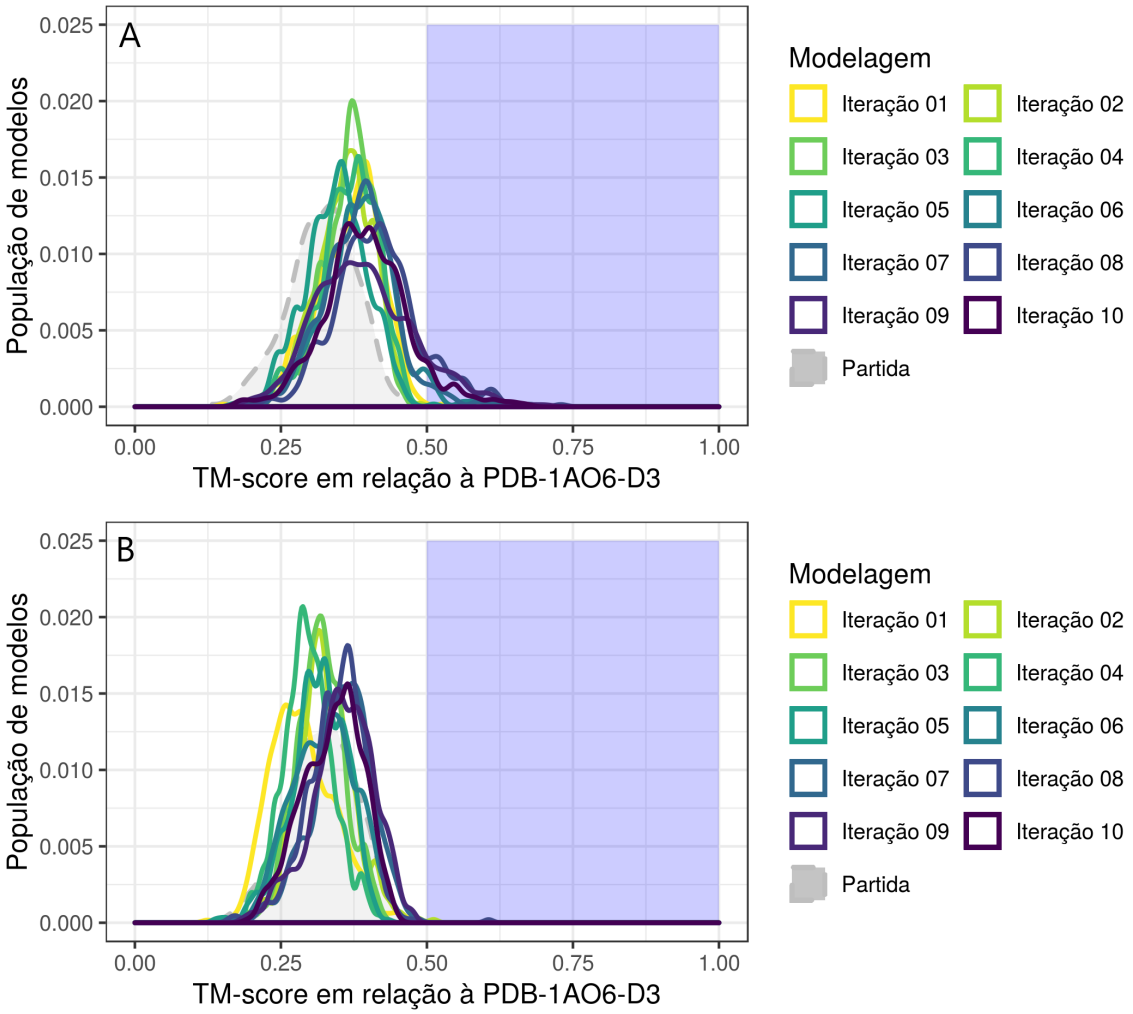

Figura 4.14: Progresso da qualidade das modelagens, na reexecução do experimento no sistema HSA-D3, com recuperação de 34 restrições e ao longo de dez iterações, para os quatro critérios já experimentados: coeficiente bisserial empregando a similaridade ao melhor modelo (A) e coeficiente bisserial empregando a similaridade ao modelo de maior ProQ3D-tmscore (B)

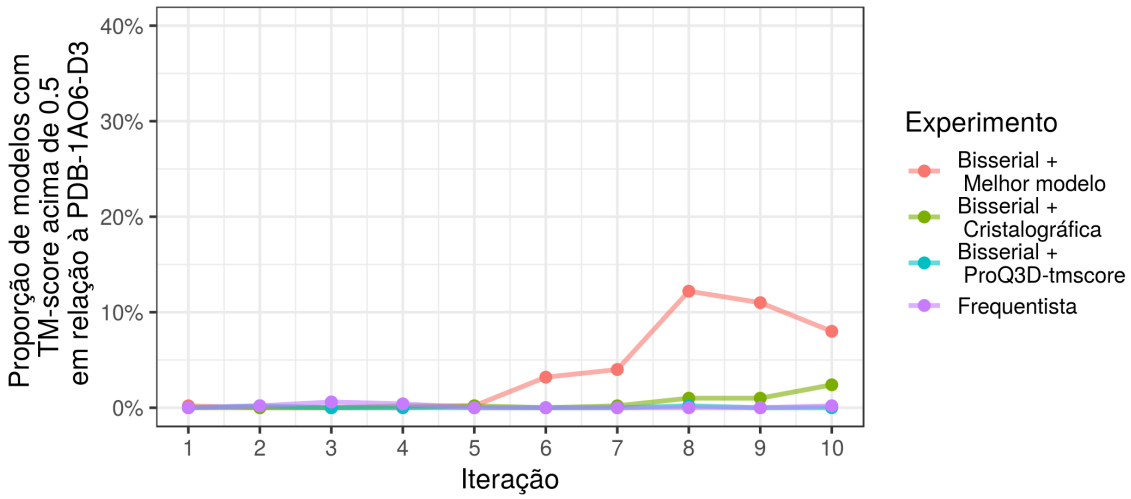

Figura 4.15: Progresso da qualidade das modelagens, na reexecução do experimento no sistema HSA-D3, com recuperação de 34 restrições, para os quatro critérios já experimentados e ao longo de dez iterações.

Tabela 4.3: Figuras de mérito para execução original e reexecução da modelagem usando como critério de recuperação o coeficiente bisserial e o melhor modelo

Experimento	Iteração	Conjunto Recuperado		Figuras de Mérito	
		Tamanho do conjunto recuperado	Restrições cristalográficas recuperadas	Taxa de Verdadeiros Positivos*	Taxa de Falsos Positivos*
Original	1	58	23	0,397	0,603
	2	58	26	0,448	0,552
	3	58	25	0,431	0,569
	4	58	21	0,362	0,638
	5	58	23	0,397	0,603
	6	58	20	0,345	0,655
	7	58	23	0,397	0,603
	8	58	25	0,431	0,569
	9	58	21	0,362	0,638
	10	58	24	0,414	0,586
Reexecução	1	34	20	0,588	0,412
	2	34	19	0,559	0,441
	3	34	22	0,647	0,353
	4	34	15	0,441	0,559
	5	34	18	0,529	0,471
	6	34	16	0,471	0,529
	7	34	24	0,706	0,294
	8	34	21	0,618	0,382
	9	34	24	0,706	0,294
	10	34	20	0,588	0,412

* - considerou-se verdadeiro positivo qualquer restrição cristalográfica.

Tabela 4.4: Figuras de mérito para execução original e reexecução da modelagem usando como critério de recuperação o coeficiente bisserial e a estrutura cristalográfica

Experimento	Iteração	Conjunto Recuperado		Figuras de Mérito	
		Tamanho do conjunto recuperado	Restrições cristalográficas recuperadas	Taxa de Verdadeiros Positivos*	Taxa de Falsos Positivos*
Original	1	58	23	0,397	0,603
	1	58	26	0,448	0,552
	2	58	22	0,379	0,621
	3	58	21	0,362	0,638
	4	58	20	0,345	0,655
	5	58	23	0,397	0,603
	6	58	25	0,431	0,569
	7	58	25	0,431	0,569
	8	58	22	0,379	0,621
	9	58	24	0,414	0,586
	10	58	26	0,448	0,552
Reexecução	1	58	23	0,397	0,603
	1	34	17	0,500	0,500
	2	34	19	0,559	0,441
	3	34	21	0,618	0,382
	4	34	18	0,529	0,471
	5	34	17	0,500	0,500
	6	34	19	0,559	0,441
	7	34	22	0,647	0,353
	8	34	20	0,588	0,412
	9	34	21	0,618	0,382
	10	34	22	0,647	0,353

* - considerou-se verdadeiro positivo qualquer restrição cristalográfica.

Capítulo 5

Considerações Finais

A modelagem biomolecular é uma área de muito interesse para diversas áreas do conhecimento científico à proporção que ela permite elucidar estruturas tridimensionais de maneira complementar a técnicas experimentais de investigação da estrutura de proteínas. Modelar proteínas para as quais existem poucos homólogos de estrutura conhecida é o principal desafio nessa área, e também aquele cuja urgência de desenvolvimento é a mais preponderante. No caso da modelagem *de novo* de proteínas por meio de abordagens *ab initio* baseadas em conhecimento, um dos principais problemas é lidar com a propagação de erros dos campos de força clássicos, que se torna muito impeditiva no caso de proteínas com mais de 100 aminoácidos na sua sequência. Embora alguns cientistas tentem sobrepôr esse problema por meio de técnicas que aumentam a precisão dos campos de força, outra maneira de superar esse empecilho é partir para estratégias de modelagem assistida, na qual dados experimentais (principalmente espectroscópicos) sobre a proteínas em solução são utilizados para customizar os campos de força, agregando aos termos comuns e abrangentes informações específicas do sistema.

Dentre as muitas estratégias utilizadas para gerar dados que auxilia modelagens, a Espectrometria de Massas de Cross-Linking é uma técnica que surge em meados dos anos 2000, resgatando protocolos de mais de 40 anos de conexão de resíduos em proteínas mas revolucionando a obtenção de dados com espectrômetros de massas e ferramentas poderosas de interpretação e anotação de espectros. Ainda assim, por conta de diversos fatores - entre eles a liberdade conformacional da proteína em solução e os protocolos de tratamento de dados e anotação espectral -, muitos dos *cross-links* observados são falsos positivos em relação às estruturas cristalográficas de proteínas conhecidas submetidas a essas técnicas. Portanto, se faz

necessário criar uma estratégia que contribua para a seleção desses dados, uma vez que a curadoria artesanal é custosa do ponto de vista humano.

Em contraste com técnicas baseadas em frequência, sugeriu-se nesse trabalho o resgate de uma técnica clássica da psicometria, denominada coeficiente de correlação ponto-bisserial, que é uma medida de discriminância entre uma variável discreta (ou binária) e outra, contínua. Sugeriu-se que, empregando esse coeficiente aliado com uma variável contínua representativa, e um processo iterativo onde, a cada passo, o conjunto de restrições recuperado fosse reavaliado, essa abordagem superaria em desempenho a recuperação frequentista.

Para tal, diversos testes preliminares foram feitos e, por fim, chegou-se à conclusão de que um possível protocolo envolveria eleger um modelo de referência a cada modelagem e tentar recuperar restrições que melhro discriminassem modelos mais e menos similares a esse modelo eleito. Essa abordagem foi testada de três maneiras diferentes, a saber: utilizando sempre como referência a estrutura cristalográfica, utilizando como referência o modelo mais similar à estrutura cristalográfica e, por fim, utilizando como melhor modelo aquele selecionado por um classificador independente, chamado PROQ3D, que tem obtido bons resultados na competição CASP.

Quatro sistemas proteicos foram avaliados, sendo eles a proteína SALBIII e os três domínios da proteína HSA.

Modelagens da proteína SALBIII

A proteína SALBIII comportou-se de maneira esperada, na qual a recuperação frequentista, mesmo após 10 iterações, não foi capaz de gerar incremente substancial na qualidade da modelagem, gerando conjuntos em que menos de 1% das proteínas de fato alcançavam TM-Score de 0,5 em relação à estrutura cristalgoraáfica, em detrimento das estratégias baseadas no coeficiente bisserial, onde aquelas empregando como referência a estrutura cristalográfica e o melhor modelo, a cada iteração, obtiveram os melhores desempenhos - respectivamente, 23% e 21% de bons modelos - e a estrutura eleita pelo PROQ3D obteve desempenho intermediário, chegando a 10% de modelos adequados.

Nesses passos de modelagem, a inesperada instabilidade na tonicidade da qualidade geral levou à conclusão de que os principais fatores que afetam a qualidade tanto de cada iteração quanto

do processo como um todo são o tamanho do conjunto de restrições recuperado, a qualidade do modelo selecionado em cada iteração e a concordância do viés introduzido pelos verdadeiros positivos em comparação aos falsos positivos do conjunto recuperado.

Modelagens da proteína HSA

A proteína HSA apresentou resultados diferentes para cada um dos três modelos, o que já era esperado porque, a despeito de serem estruturalmente similares, a reatividade de cada sequência e a qualidade dos dados disponíveis era diferente para cada um deles.

Em todos os casos, a recuperação frequentista fracassa novamente, de maneira similar à SALBIII.

No caso da HSA-D1 e HSA-D2, as estratégias que empregavam o coeficiente bisserial em referência à estrutura cristalográfica e à melhor estrutura obtiveram, sim, bom desempenho: no caso da HSA-D1, alcançando respectivamente e, no caso da HSA-D2, chegando a 33% e 22%, também respectivamente. Para esses sistemas, a seleção baseada no PROQ3D infelizmente não obteve resultados superiores à seleção frequentista, o que mostra a importância da seleção de um bom modelo como referência.

No caso da HSA-D3, nenhuma das estratégias foi capaz de gerar quantidade significativa de bons modelos. Nesse caso, a causa do fracasso está mais relacionada ao tamanho, taxa de falsos positivos e discordância geral do conjunto de restrições, outro fator determinante para o avanço das modelagens.

5.1 Perspectivas

A expansão do protocolo desenvolvido para outros sistemas pode ser interessante do ponto de vista de investigar o desempenho da recuperação de restrições em mais sistemas proteicos de características diferentes. Seria muito interessante comparar, por exemplo, o desempenho entre proteínas, α, β e também $\alpha\beta$. Disso depende, no entanto, a coleta de dados instrumentais, que é um custo e o principal gargalo do projeto.

Além disso, existem ainda algumas sugestões de como melhorar esse protocolo, como,

por exemplo, utilizar medidas de agrupamento para aferir a convergência da modelagem, descartando um passo iterativo no caso de aumento da dispersão de modelos, ou realizando experimentos de *bootstrapping* em modelagens paralelas de uma mesma iteração com conjuntos diferentes, paras selecionar aquele que será considerado para recuperação de restrições. No entanto, a maioria das sugestões requer ou abordagens combinatórias ou experimentos com grande custo computacional.

Certamente um dos pontos principais que deve ser explorado é o tamanho do conjunto de restrições a ser recuperado. De fato, é sensível que ele foi, no caso dos quatro sistemas, a principal fonte de falhas. No entanto, esse ponto é extremamente sensível, porque depende não só do tamanho da sequência primária como foi sugerido aqui para testes preliminares, mas também da quantidade e qualidade dos dados experimentais obtidos. Há experimentos de XL-MS, inclusive na competição CASP, para os quais foram obtidas apenas 10 restrições para uma proteína de 200 aminoácidos. Esse certamente é o tipo de resposta que virá de uma interação muito forte com os expecialistas nos experimentos de espectrometria de massas e do melhoramento da própria química dos *cross-linkers*.

Em relação á transferência de conhecimento, todo o protocolo de recuperação de restrições foi escrito e disponibilizado na versão alpha do pacote de softwares ZedXL, que já conta com uma versão em linha de comando para execução automatizada do protocolo em clusters de computação, em Linux e também em Windows e Macintosh. O pacote, de código aberto, pode ser acessado em https://github.com/Hugemiler/ZedXL, para uso de toda a comunidade científica. Num futuro próximo, será elaborada a documentação completa do pacote dentro das diretrizes da iniciativa global Bioconductor de pacotes estatísticos para bioinformática, o que potencializa a visibilidade dos softwares desenvolvidos e prepara todo o pacote para publicação da versão definitiva, que será acompanhada da elaboração de um artigo de metodologia.

Além disso, futuramente dentro do CCES, será estabelecido um servidor com todo o fluxo de trabalho automático e uma API padrão RESTful que disponibilizará a metodologia desenvolvida à comunidade científica na forma de serviço (*PaaS – Platform as a Service*).

Bibliografia

[1] David L Nelson e Michael M Cox. *Lehninger Principles of Biochemistry, Fourth Edition*. Ed. por Freeman. Fourth Edition. 2004.

[2] D. Voet e J.G. Voet. *Biochemistry, 4th Edition*. John Wiley & Sons, 2010. ISBN: 9781118139936. URL: https://books.google.com.br/books?id=ne0bAAAAQBAJ.

[3] Lukasz Slabinski et al. "The challenge of protein structure determination–lessons from structural genomics." Em: *Protein science : a publication of the Protein Society* 16.11 (2007), pp. 2472–82. ISSN: 0961-8368. DOI: 10.1110/ps.073037907.

[4] Ken A. Dill et al. "The Protein Folding Problem". Em: *Annual Review of Biophysics* 37.1 (2008), pp. 289–316. ISSN: 1936-122X. DOI: 10.1146/annurev.biophys.37.092707.153558.

[5] Christopher D. Putnam et al. "X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution". Em: *Quarterly Reviews of Biophysics* 40.03 (2007), pp. 191–285. ISSN: 0033-5835. DOI: 10.1017/S0033583507004635.

[6] Jin Seob Kim, Bijan Afsari e Gregory S Chirikjian. "Cross-Validation of Data Compatibility Between Small Angle X-ray Scattering and Cryo-Electron Microscopy." Em: *Journal of computational biology : a journal of computational molecular cell biology* 24.1 (2017), pp. 13–30. ISSN: 1557-8666. DOI: 10.1089/cmb.2016.0139.

[7] Alexander Grishaev et al. "Refinement of Multidomain Protein Structures by Combination of Solution Small-Angle X-ray Scattering and NMR Data". Em: *Journal of the American Chemical Society* 127.47 (2005), pp. 16621–16628. DOI: 10.1021/ja054342m.

[8] Cyrus Levinthal. "How to fold graciously". Em: *Mossbauer spectroscopy in biological systems* 67 (1969), pp. 22–24.

[9] Cyrus Levinthal. "Are there pathways for protein folding?" Em: *Journal de Chimie Physique* 65 (1968), pp. 44–45. ISSN: 0021-7689. DOI: 10.1051/jcp/1968650044.

[10] Christian B. Anfinsen. "Principles that govern the folding of protein chains." Em: *Science (New York, N.Y.)* 181.4096 (1973), pp. 223–30. ISSN: 0036-8075. DOI: 10.1126/SCIENCE.181.4096.223.

[11] Martin Karplus. "Protein folding: theoretical studies of thermodynamics and dynamics". Em: *Protein folding* (1992), pp. 127–196.

[12] Leandro Martínez. "Introducing the Levinthal's Protein Folding Paradox and Its Solution". Em: *Journal of Chemical Education* 91 (nov. de 2014), pp. 1918–1923. DOI: 10.1021/ed300302h.

[13] Michael Levitt e Arieh Warshel. "Computer simulation of protein folding". Em: *Nature* 253.5494 (1975), pp. 694–698. ISSN: 0028-0836. DOI: 10.1038/253694a0.

[14] W. F. van Gunsteren e Berendsen. *Groningen Molecular Simulation (GROMOS) Library Manual*. Biomos, Nijenborgh 16, Groningen, NL. 1987.

[15] Bernard R. Brooks et al. "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations". Em: *Journal of Computational Chemistry* 4.2 (1983), pp. 187–217. DOI: 10.1002/jcc.540040211.

[16] William L. Jorgensen e Julian Tirado-Rives. "The OPLS potential functions for proteins. Energy minimizations for crystals of cyclic peptides and crambin". Em: *Journal of The American Chemical Society - J AM CHEM SOC* 110 (1988), pp. 1657–1666. DOI: 10.1021/ja00214a001.

[17] Peter L; Freddolino et al. "Challenges in protein folding simulations: Timescale, representation, and analysis." Em: *Nature physics* 6.10 (2010), pp. 751–758. ISSN: 1745-2473. DOI: 10.1038/nphys1713.

[18] Guha Jayachandran, V Vishal e Vijay S. Pande. "Using massively parallel simulation and Markovian models to study protein folding: Examining the dynamics of the villin headpiece". Em: *The Journal of Chemical Physics* 124.16 (2006), p. 164902. DOI: 10.1063/1.2186317. URL: https://doi.org/10.1063/1.2186317.

[19] David E. Kim et al. "Sampling Bottlenecks in De novo Protein Structure Prediction". Em: *Journal of Molecular Biology* 393.1 (2009), pp. 249–260. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2009.07.063.

[20] Kevin A. Dill. "Additivity principles in biochemistry." Em: *The Journal of biological chemistry* 272.2 (1997), pp. 701–4. ISSN: 0021-9258. DOI: 10.1074/JBC.272.2.701.

[21] John C. Faver et al. "The Energy Computation Paradox and ab initio Protein Folding". Em: *PLoS ONE* 6.4 (2011). Ed. por Collin M. Stultz, e18868. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0018868.

[22] Keehyoung Joo et al. "Data-assisted protein structure modeling by global optimization in CASP12". Em: *Proteins: Structure, Function, and Bioinformatics* 86.S1 (2018), pp. 240–246. DOI: 10.1002/prot.25457.

[23] Daniel Russel et al. "Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies". Em: *PLOS Biology* 10.1 (2012), pp. 1–5. DOI: 10.1371/journal.pbio.1001244.

[24] Andrea Sinz. "Cross-Linking/Mass Spectrometry for Studying Protein Structures and Protein–Protein Interactions: Where Are We Now and Where Should We Go from Here?" Em: *Angewandte Chemie International Edition* 57.22 (2018), pp. 6390–6396. DOI: 10.1002/anie.201709559.

[25] Christopher Clegg e Donal Hayes. "Identification of Neighbouring Proteins in the Ribosomes of Escherichia coli". Em: *European Journal of Biochemistry* 42.1 (1974), pp. 21–28. DOI: 10.1111/j.1432-1033.1974.tb03309.x.

[26] J Rappsilber et al. "A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass spectrometry." Em: *Analytical chemistry* 72.2 (2000), pp. 267–75. ISSN: 0003-2700. DOI: 10.1021/ac991081o.

[27] Michal Sharon et al. "Structural Organization of the 19S Proteasome Lid: Insights from MS of Intact Complexes". Em: *PLOS Biology* 4.8 (2006). DOI: 10.1371/journal.pbio.0040267.

[28] Jochen Walz et al. "26S proteasome structure revealed by three-dimensional electron microscopy". Em: *Journal of structural biology* 121.1 (1998), pp. 19–29. DOI: 10.1006/jsbi.1998.3958.

[29] Andrea Sinz. "Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein–protein interactions". Em: *Mass Spectrometry Reviews* 25.4 (2006), pp. 663–682. DOI: 10.1002/mas.20082.

[30] Claudio Ciferri et al. "Implications for kinetochore-microtubule attachment from the structure of an engineered Ndc80 complex". Em: *Cell* 133.3 (2008), pp. 427–439. DOI: 10.1016/j.cell.2008.03.020.

[31] Alessio Maiolica et al. "Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching". Em: *Molecular & Cellular Proteomics* 6.12 (2007), pp. 2200–2211. DOI: 10.1074/mcp.M700274-MCP200.

[32] Oliver Rinner et al. "Identification of cross-linked peptides from large sequence databases". Em: *Nature methods* 5.4 (2008), p. 315. DOI: 10.1038/nmeth.1192.

[33] Zhuo Angel Chen et al. "Architecture of the RNA polymerase II–TFIIF complex revealed by cross-linking and mass spectrometry". Em: *The EMBO Journal* 29.4 (2010), pp. 717–726. ISSN: 0261-4189. DOI: 10.1038/emboj.2009.401.

[34] Mariana Fioramonte et al. "XPlex: An Effective, Multiplex Cross-Linking Chemistry for Acidic Residues". Em: *Analytical Chemistry* 90.10 (2018). PMID: 29565564, pp. 6043–6050. DOI: 10.1021/acs.analchem.7b05135.

[35] Andrew N. Holding. "XL-MS: Protein cross-linking coupled with mass spectrometry". Em: *Methods* 89 (2015), pp. 54–63. ISSN: 10462023. DOI: 10.1016/j.ymeth.2015.06.010.

[36] Lutz Fischer e Juri Rappsilber. "Quirks of Error Estimation in Cross-Linking/Mass Spectrometry". Em: *Analytical Chemistry* 89.7 (2017). ISSN: 15206882. DOI: 10.1021/acs.analchem.6b03745.

[37] Diogo B Lima et al. "SIM-XL: A powerful and user-friendly tool for peptide cross-linking analysis". Em: *Journal of proteomics* 129 (2015), pp. 51–55. ISSN: 1874-3919. DOI: 10.1016/j.jprot.2015.01.013.

[38] Clinton Yu et al. "Developing a Multiplexed Quantitative Cross-Linking Mass Spectrometry Platform for Comparative Structural Analysis of Protein Complexes". Em: *Analytical Chemistry* 88.20 (2016). ISSN: 15206882. DOI: 10.1021/acs.analchem.6b03148.

[39] Fan Liu et al. "Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification". Em: *Nature Communications* 8 (2017). ISSN: 2041-1723. DOI: 10.1038/ncomms15473.

[40] Marie E. Yurkovich et al. "A Late-Stage Intermediate in Salinomycin Biosynthesis Is Revealed by Specific Mutation in the Biosynthetic Gene Cluster". Em: *ChemBioChem* 13.1 (2012), pp. 66–71. ISSN: 14394227. DOI: 10.1002/cbic.201100590.

[41] Hanna Luhavaya et al. "Enzymology of Pyran Ring A Formation in Salinomycin Biosynthesis." Em: *Angew.Chem.Int.Ed.Engl.* 54 (2015), pp. 13622–13625. DOI: 10.2210/PDB5CXO/PDB.

[42] Ricardo N dos Santos et al. "Enhancing protein fold determination by exploring the complementary information of chemical cross-linking and coevolutionary signals". Em: *Bioinformatics* 34.13 (2018), pp. 2201–2208. DOI: 10.1093/bioinformatics/bty074.

[43] Allan J R Ferrari, Fabio C Gozzo e Leandro Martínez. "Statistical force-field for structural modeling using chemical cross-linking/mass spectrometry distance constraints". Em: *Bioinformatics* (2019). DOI: 10.1093/bioinformatics/btz013.

[44] D C Carter e J X Ho. "Structure of serum albumin." Em: *Advances in protein chemistry* 45 (1994), pp. 153–203. ISSN: 0065-3233. URL: http://www.ncbi.nlm.nih.gov/pubmed/8154369.

[45] M Dockal, D C Carter e F Rüker. "The three recombinant domains of human serum albumin. Structural characterization and ligand binding properties." Em: *The Journal of biological chemistry* 274.41 (1999), pp. 29303–10. DOI: 10.1074/JBC.274.41.29303.

[46] S. Sugio et al. "Crystal structure of human serum albumin at 2.5 A resolution." Em: *Protein Eng.* 12 (1999), pp. 439–446. DOI: 10.2210/PDB1AO6/PDB.

[47] T Peters. "Serum albumin." Em: *Advances in protein chemistry* 37 (1985), pp. 161–245. ISSN: 0065-3233. URL: http://www.ncbi.nlm.nih.gov/pubmed/3904348.

[48] R G Reed et al. "Fragments of bovine serum albumin produced by limited proteolysis. Conformation and ligand binding." Em: *Biochemistry* 14.21 (1975), pp. 4578–83. ISSN: 0006-2960.

[49] D J Ledden, R C Feldhoff e S K Chan. "Characterization of fragments of human albumin purified by Cibacron blue F3GA affinity chromatography." Em: *The Biochemical journal* 205.2 (1982), pp. 331–7. ISSN: 0264-6021.

[50] O J Bos et al. "Drug-binding and other physicochemical properties of a large tryptic and a large peptic fragment of human serum albumin." Em: *Biochimica et biophysica acta* 953.1 (1988), pp. 37–47. ISSN: 0006-3002.

[51] Kim T. Simons et al. "Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions". Em: *Journal of Molecular Biology* 268.1 (1997), pp. 209–225. ISSN: 0022-2836. DOI: https://doi.org/10.1006/jmbi.1997.0959.

[52] Kim T. Simons et al. "Ab initio protein structure prediction of CASP III targets using ROSETTA". Em: *Proteins: Structure, Function, and Bioinformatics* 37.S3 (1999), pp. 171–176. DOI: 10.1002/(SICI)1097-0134(1999)37:3+<171::AID-PROT21>3.0.CO;2-Z.

[53] Brian Kuhlman et al. "Design of a Novel Globular Protein Fold with Atomic-Level Accuracy". Em: *Science* 302 (2003), pp. 1364–1368. DOI: 10.2210/PDB1QYS/PDB. URL: https://www.rcsb.org/structure/1QYS.

[54] RosettaCommons. *Full Atom Representation vs Centroid Representation*. URL: https://www.rosettacommons.org/demos/latest/tutorials/full%7B%5C_%7Datom%7B%5C_%7Dvs%7B%5C_%7Dcentroid/fullatom%7B%5C_%7Dcentroid%7B%5C#%7Dexample (acesso em 30/11/2018).

[55] Kim T. Simons et al. "Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins". Em: *Proteins: Structure, Function, and Bioinformatics* 34.1 (1999), pp. 82–95. DOI: 10.1002/(SICI)1097-0134(19990101)34:1<82::AID-PROT7>3.0.CO;2-A.

[56] RosettaCommons. *Score Function History*. URL: https://www.rosettacommons.org/docs/latest/rosetta%5C_basics/scoring/Scorefunction-History (acesso em 30/11/2018).

[57] Andrew Leaver-Fay et al. "Chapter Six - Scientific Benchmarks for Guiding Macromolecular Energy Function Improvement". Em: *Methods in Protein Design*. Ed. por Amy E. Keating. Vol. 523. Methods in Enzymology. Academic Press, 2013, pp. 109–143. DOI: 10.1016/B978-0-12-394292-0.00006-0.

[58] Rebecca F. Alford et al. "The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design". Em: *Journal of Chemical Theory and Computation* 13.6 (2017). PMID: 28430426, pp. 3031–3048. DOI: 10.1021/acs.jctc.7b00125.

[59] Abdullah Kahraman et al. "Cross-Link Guided Molecular Modeling with ROSETTA". Em: *PLoS ONE* 8.9 (2013). Ed. por Narcis Fernandez-Fuentes, e73411. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0073411.

[60] Stefan Kalkhof et al. "Computational modeling of laminin N-terminal domains using sparse distance constraints from disulfide bonds and chemical cross-linking". Em: *Proteins: Structure, Function, and Bioinformatics* 78.16 (2010), pp. 3409–3427. DOI: 10.1002/prot.22848.

[61] G. V. Glass e K. D. Hopkins. *Statistical Methods in Education and Psychology*. Third Edition. Allyn & Bacon, 1995. ISBN: 0-205-14212-5.

[62] J. M. Linacre. "The Expected Value of a Point-Biserial (or Similar) Correlation." Em: *Rasch Measurement Transactions* 22.1 (2008), p. 1154. ISSN: 1051-0796.

[63] Yang Zhang e Jeffrey Skolnick. "Scoring function for automated assessment of protein structure template quality". Em: *Proteins: Structure, Function, and Bioinformatics* 57.4 (2004), pp. 702–710. DOI: 10.1002/prot.20264.

[64] Jinrui Xu e Yang Zhang. "How significant is a protein structure similarity with TM-score = 0.5?" Em: 26.7 (2010), pp. 889–895. DOI: 10.1093/bioinformatics/btq066.

[65] Helen M. Berman et al. "The Protein Data Bank". Em: *Nucleic Acids Research* 28.1 (2000), pp. 235–242. DOI: 10.1093/nar/28.1.235.

[66] John C. Kendrew et al. "A three-dimensional model of the myoglobin molecule obtained by x-ray analysis". Em: *Nature* 181.4610 (1958), pp. 662–666. DOI: 10.1038/181662a0.

[67] Yigong Shi. "A glimpse of structural biology through X-ray crystallography." Em: *Cell* 159.5 (2014), pp. 995–1014. ISSN: 1097-4172. DOI: 10.1016/j.cell.2014.10.051.

[68] K. Ravi Acharya e Matthew D. Lloyd. "The advantages and limitations of protein crystal structures". Em: *Trends in Pharmacological Sciences* 26.1 (2005), pp. 10–14. ISSN: 0165-6147. DOI: https://doi.org/10.1016/j.tips.2004.10.011.

[69] H. Poincaré. *The Value of Science*. Cosimo Classics. Science. Cosimo Classics, 2007. ISBN: 9781602065048. URL: https://books.google.com.br/books?id=%5C_kHdT7U77dcC.

[70] Andriy Kryshtafovych et al. "Assessment of model accuracy estimations in CASP12". Em: *Proteins: Structure, Function, and Bioinformatics* 86.S1 (2018), pp. 345–360. DOI: 10.1002/prot.25371.

[71] Arne Elofsson et al. "Methods for estimation of model accuracy in CASP12". Em: *Proteins: Structure, Function, and Bioinformatics* 86.S1 (2018), pp. 361–373. DOI: 10.1002/prot.25395.

[72] Björn Wallner e Arne Elofsson. "Can correct protein models be identified?" Em: *Protein science : a publication of the Protein Society* 12.5 (2003), pp. 1073–86. ISSN: 0961-8368. DOI: 10.1110/ps.0236803.

[73] Arjun Ray, Erik Lindahl e Björn Wallner. "Improved model quality assessment using ProQ2". Em: *BMC Bioinformatics* 13.1 (2012), p. 224. ISSN: 1471-2105. DOI: 10.1186/1471-2105-13-224.

[74] Karolis Uziela et al. "ProQ3: Improved model quality assessments using Rosetta energy terms". Em: *Scientific Reports* 6.1 (2016), p. 33509. ISSN: 2045-2322. DOI: 10.1038/srep33509.

[75] Karolis Uziela et al. "ProQ3D: improved model quality assessments using deep learning". Em: *Bioinformatics* 33.10 (2017), pp. 1578–1580. DOI: 10.1093/bioinformatics/btw819.

[76] Eduardo Tejada, Rosane Minghim e Luis Gustavo Nonato. "On improved projection techniques to support visual exploration of multi-dimensional data sets". Em: *Information Visualization* 2 (2003), pp. 218–231. DOI: 10.1057/palgrave.ivs.9500054.

[77] Antonio B. Oliveira et al. "Visualization of Protein Folding Funnels in Lattice Models". Em: *PLoS ONE* 9.7 (2014). Ed. por Yaakov Koby Levy, e100861. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0100861.

[78] Leandro Martinez, Allan Ferrari e Fabio Cesar Gozzo. "TopoLink: A software to validate structural models using chemical crosslinking constraints". Em: (2017). DOI: 10.1038/protex.2017.035.

[79] Stephen F. Altschul et al. "Basic local alignment search tool". Em: *Journal of Molecular Biology* 215.3 (1990), pp. 403–410. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(05)80360-2.

[80] B. E. Suzek et al. "UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches". Em: *Bioinformatics* 31.6 (2015), pp. 926–932. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu739.

[81] David T. Jones. "Protein secondary structure prediction based on position-specific scoring matrices11Edited by G. Von Heijne". Em: *Journal of Molecular Biology* 292.2 (1999), pp. 195–202. ISSN: 0022-2836. DOI: 10.1006/jmbi.1999.3091.

[82] Gianluca Pollastri et al. "Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles". Em: *Proteins: Structure, Function, and Bioinformatics* 47.2 (2002), pp. 228–235. DOI: 10.1002/prot.10082.

[83] Leandro Martínez, Roberto Andreani e José Mario Martínez. "Convergent algorithms for protein structural alignment". Em: *BMC Bioinformatics* 8.1 (2007), p. 306. ISSN: 1471-2105. DOI: 10.1186/1471-2105-8-306.

[84] Leandro Martínez, Allan Ferrari e Fabio C. Gozzo. *A free-energy inspired score for the evaluation of structural models*. URL: http://leandro.iqm.unicamp.br/gscore/home.shtml (acesso em 30/11/2018).

[85] R Development Core Team. *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2008. URL: http://www.R-project.org.

[86] David E. Kim et al. "One contact for every twelve residues allows robust and accurate topology-level protein structure modeling". Em: *Proteins: Structure, Function, and Bioinformatics* 82.S2 (2014), pp. 208–218. DOI: 10.1002/prot.24374.

[87] Varun Mandalaparthy et al. "Exploring the effects of sparse restraints on protein structure prediction". Em: *Proteins: Structure, Function, and Bioinformatics* 86.2 (2018), pp. 248–262. ISSN: 08873585. DOI: 10.1002/prot.25438.