**UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS**

Samuel Reghim Silva

**Representações de superfícies moleculares em
harmônicos esféricos para simulação de formação de
complexos entre proteínas**

São Carlos

2018

Samuel Reghim Silva

**Representações de superfícies moleculares em
harmônicos esféricos para simulação de formação de
complexos entre proteínas**

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

Área de concentração: Física Aplicada
Opção: Física Computacional

Orientador: Prof. Dr. Glaucius Oliva
Coorientador: Dr. Rinaldo Wander Montalvão

Versão corrigida
(Versão original disponível na Unidade que aloja o Programa)

São Carlos
2018

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

Silva, Samuel Reghim

Representações de superfícies moleculares em harmônicos esféricos para simulação de formação de complexos entre proteínas / Samuel Reghim Silva; orientador Glaucius Oliva; co-orientador Rinaldo Wander Montalvão - versão corrigida -- São Carlos, 2018.

123 p.

Tese (Doutorado - Programa de Pós-Graduação em Física Aplicada Computacional) -- Instituto de Física de São Carlos, Universidade de São Paulo, 2018.

1. Reconstrução de imagens. 2. Superfície molecular. 3. Atracamento de proteínas. 4. Simulação. I. Oliva, Glaucius, orient. II. Montalvão, Rinaldo Wander, co-orient. III. Título.

Dedico esta tese aos meus pais, João e Sueli, e à minha irmã, Carolina.

AGRADECIMENTOS

Aos meus pais João Rodrigues da Silva e Sueli dos Santos Silva e à minha irmã favorita, Carolina Reghim Silva, pelo apoio incondicional, carinho, paciência, motivação, compreensão da distância e exemplo de persistência.

À querida Marília Vilela pelo companheirismo, carinho, paciência e motivação.

Ao meu orientador, Dr. Rinaldo Wander Montalvão, pela confiança e oportunidade de pesquisa, pelo treinamento técnico-científico, pelos ensinamentos humanos, pela motivação e pelo exemplo de excelência.

Ao Professor Glaucius Oliva pela confiança e disponibilidade de assumir a orientação, oferecendo imprescindível ajuda e possibilitando a conclusão deste projeto.

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), pelo financiamento deste projeto, através dos processos de números 2013/20929-4 e 2016/20366-8, sem o qual sua realização teria sido ainda mais difícil.

Aos Professores Hélio Guardia e Ana Paula Araújo pela inestimável ajuda e orientação, e por não pouparem esforços em auxiliar minha transição para a Universidade de São Paulo.

Ao Professor Michele Vendruscolo pela oportunidade de pesquisa no Centre for Misfolding Diseases do Department of Chemistry, University of Cambridge.

Aos colegas do Department of Chemistry pelo acolhimento e respeito.

Aos amigos Antônio “Tonho” Marinho e Katarina Holanda, Evandro “Evandrinho” Ares, Fernando Caires, Ivan Silva, Leonardo Almeida, Alain Mera, Sabrina Matos, Amanda “Amandinha” Câmara, Fernando Molina e Luciana Midori, Rodrigo Berté, Fernanda Sala, Tatiana Watanabe e Patrícia Pantoja pelo excelente companheirismo nos altos e baixos desta longa jornada de amadurecimento técnico e humano, estimulando muitas risadas, debates intelectuais, aventuras e noites em claro de *videogames* e de trabalho.

Aos colegas de república Cristian, Fábio, Tiago, Eduar e Yago pela companhia.

À Sra. Cherry Topden e à sua família pela hospedagem, ajuda e pela querida companhia.

A Richard Morris e Ligang Liu, pelas discussões acerca de representação de funções em Zernike 3D.

A Daniele Panozo pelo auxílio com mapeamentos esféricos.

A Holger Gohlke pelo código-fonte de seu programa de reconstrução de funções em Zernike 3D.

Aos Professores João Renato Muniz, Luciano Fontoura e Luis Gustavo Nonato pelos comentários enriquecedores sobre este projeto.

Às bibliotecárias do Campus São Carlos da Universidade de São Paulo pelo desenvolvimento da classe $\text{\LaTeX}$, e às do Instituto de Física de São Carlos pela seriedade e atenção na revisão desta tese.

Aos técnicos do ScInfor pela agilidade e seriedade na prestação de serviços de informática, aos do ScGraf pelas impressões em alta qualidade e curto prazo, e a todos os demais funcionários da Universidade de São Paulo pelo zelo com que mantêm este centro de excelência através das equipes de limpeza, manutenção, vigilância, técnica e administrativa.

A todos os desenvolvedores de software livre pela criação de incríveis ferramentas, especialmente aos dos projetos GNU, Linux, Geany, GCC, Meshlab, Matplotlib, $\text{\TeX}$ e $\text{\LaTeX}$.

À Universidade de São Paulo e ao Instituto de Física de São Carlos pela oportunidade de excelente formação intelectual e infra-estrutura de alto nível, com destaque para a Biblioteca do Instituto e para o Restaurante Universitário.

“É fazendo que se aprende a fazer aquilo que se deve aprender a fazer.”

Aristóteles

RESUMO

SILVA, S. R. **Representações de superfícies moleculares em harmônicos esféricos para simulação de formação de complexos entre proteínas.** 2018. 123p. Tese (Doutorado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2018.

O uso de programas de computador para simular a formação de complexos entre proteínas é uma abordagem importante para melhor compreensão de como estas moléculas interagem. A representação paramétrica e a representação em polinômios tridimensionais de Zernike são ambas descrições compactas de superfícies moleculares baseadas em harmônicos esféricos utilizadas para visualização e comparação de superfícies moleculares e para atracamento de proteínas. Entretanto, apresentam limitações como restrição à topologia da superfície e dificuldade de representação de funções arbitrárias. Neste estudo, procurou-se refinar a capacidade de representação destes métodos para obtenção de elevada qualidade de reprodução e aplicabilidade a superfícies de topologia arbitrária. Através da análise de diversos algoritmos de suas etapas, foi possível identificar os estágios de cálculo de malha triangular de superfície molecular e de mapeamento esférico como os mais influentes na qualidade da representação paramétrica em harmônicos esféricos, e a alta sensibilidade a mudança de valores nas funções projetadas na qualidade da representação em Zernike 3D. A incorporação de um método de cálculo de superfícies que gera uma malha com elevada regularidade, aliado a um moderno algoritmo de mapeamento esférico garantiu baixo nível de distorções e obtenção de superfícies reconstruídas de alta qualidade. Uma técnica de detecção de similaridade entre superfícies de diferentes conformações de um *ensemble* permitiu compartilhamento de partes da descrição entre várias superfícies, com correspondente redução de volume de dados e de demanda por processamento, e com influência controlável nas distorções introduzidas. Um modo diferente de organização dos dados de entrada causou melhoria na qualidade de reconstrução de funções gerais em Zernike 3D, embora com a introdução de um mapa para restauração da função. Os resultados obtidos indicam aplicação promissora dos métodos em *docking* de proteínas com alto nível de detalhes.

Palavras-chave: Reconstrução de imagens. Superfície molecular. Atracamento de proteínas. Simulação.

ABSTRACT

SILVA, S. R. **Representations of molecular surfaces in spherical harmonics for simulation of protein-protein complexes formation.** 2018. 123p. Tese (Doutorado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2018.

Using computer programs to simulate protein complex formation is an important approach for better comprehension of the interaction mechanisms of such molecules. The parametric representation and the Zernike polynomial method are both compact representations of molecular surfaces based on spherical harmonics, used for visualization and comparison of molecules and for protein-protein docking. They pose, however, limitations regarding surface topology and difficulty in representing arbitrary functions. In this study, the representation capacity of such methods were refined to attain high quality reproductions and applicability to arbitrary topologies. Throughout the analysis of several algorithms, the stages of surface mesh calculation and spherical mapping were identified as highly influential on the quality of the spherical harmonics parametric representation, while high sensibility to changes in the function values were identified as an influential factor for projections in 3D Zernike. A surface calculation method that generates a highly regular mesh was adopted and paired with a modern spherical mapping algorithm to yield reconstructions with low level of distortions and high quality surfaces. Similarity among surfaces from different structures in a conformational ensemble were detected to allow sharing of portions of the description among several surfaces, with corresponding reduction in data volume and processing demand and controllable influence of distortions. A new input data organization method improved the reconstruction quality of general functions in 3D Zernike, although introducing a map to restore the function. Results indicate promising application of the methods in highly detailed protein-protein docking.

Keywords: Image reconstruction. Molecular surface. Protein docking. Simulation.

LISTA DE ILUSTRAÇÕES

Figura 1 – Projeção em harmônicos esféricos com diferentes ordens de aproximação	27
Figura 2 – Valores calculados de $\delta_{ll'}\delta_{mm'}$ para harmônicos esféricos iguais	48
Figura 3 – Valores calculados de $\delta_{ll'}\delta_{mm'}$ para harmônicos esféricos diferentes	49
Figura 4 – Malhas calculadas pelo programa Surfmap para o complexo 1R2M	49
Figura 5 – Distribuição aleatória de pontos na superfície de um triângulo esférico	49
Figura 6 – Reconstrução de 1R2M utilizando o mapeamento de Surfmap e integração numérica	50
Figura 7 – Malhas referentes à superfície de 1R2M gerada por CGAL	51
Figura 8 – Mapeamento CALD para a superfície de 1R2M gerada por CGAL	52
Figura 9 – Reconstrução com CALD e IRF de 1R2M com 3557 vértices	53
Figura 10 – Reconstrução com CALD e IRF de 1R2M com 7114 vértices	53
Figura 11 – Mapeamento de 1R2M calculado pelo programa SPHARM-MAT	53
Figura 12 – Mapeamento de 1R2M calculado por ConformalizedMCF a partir de triangulação de CGAL	54
Figura 13 – Reconstrução de 1R2M com 21445 vértices utilizando CGAL e ConformalizedMCF	54
Figura 14 – Reconstrução de 1R2M com 50333 vértices utilizando CGAL e ConformalizedMCF	55
Figura 15 – Mapeamento e reconstrução da superfície <i>cortex</i>	55
Figura 16 – Mapeamentos de <i>armadillo</i> calculados por SPHARM-MAT e por ConformalizedMCF	56
Figura 17 – Reconstruções de <i>armadillo</i> calculadas por diferentes métodos	57
Figura 18 – Superfícies de 1R2M, 1R2M_B e de 2K5X_A calculadas por EDTSurf	57
Figura 19 – Mapeamentos de 1R2M, 1R2M_B e de 2K5X_A calculados por ConformalizedMCF	58
Figura 20 – Reconstruções das superfícies de 1R2M, 1R2M_B e de 2K5X_A	58
Figura 21 – Função coordenada original X de 1R2M e sua reconstrução X_{SH} usando o mapeamento de ConformalizedMCF	59
Figura 22 – Reconstruções de <i>fandisk</i> a partir dos mapeamentos de ConformalizedMCF e de ARAP	59
Figura 23 – Mapeamentos de <i>fandisk</i> calculados por ConformalizedMCF e por ARAP	60
Figura 24 – Mapeamentos esféricos de 1R2M e de 2K5X obtidos com BLD	61
Figura 25 – Reconstrução de 1R2M usando o mapeamento de BLD	62
Figura 26 – Reconstrução de 2K5X_A usando o mapeamento de BLD	62
Figura 27 – Função coordenada original X de 1R2M e sua reconstrução X_{SH} usando o mapeamento de BLD	63

Figura 28 – Mapeamentos AHSP das superfícies de 1R2M e de 2K5X_A	63
Figura 29 – Função coordenada original X de 1R2M e sua reconstrução X_{SH} usando o mapeamento de AHSP	64
Figura 30 – Reconstruções de 1R2M e de 2K5X_A a partir de mapeamentos AHSP	64
Figura 31 – Distorções no mapeamento ConformalizedMCF e na reconstrução de 1R2M	67
Figura 32 – Distorções no mapeamento AHSP e na reconstrução de 1R2M	68
Figura 33 – Passos para se fechar um túnel em superfície molecular	70
Figura 34 – Largura máxima de túnel e RMSD de reconstruções de estruturas de <i>Protein Docking Benchmark</i>	71
Figura 35 – Distribuição de RMSD das reconstruções de estruturas de <i>Protein Docking Benchmark</i>	71
Figura 36 – Distribuição de pontos na superfície esférica usando três algoritmos de amostragem	72
Figura 37 – Superfícies moleculares originais e reconstruções em harmônicos esféricos	73
Figura 38 – Curvaturas gaussiana e média na superfície de 1R2M	75
Figura 39 – Comparação entre curvatura gaussiana calculada por CGAL e calculada por coeficientes de harmônicos esféricos para 1R2M	76
Figura 40 – Distribuição de energia potencial eletrostática pela superfície de 1R2M	77
Figura 41 – RMSD de reconstruções de 2LJ5 com número variado de coeficientes da estrutura de referência	79
Figura 42 – Distribuição de valores para o coeficiente de maior variação segundo MAD e segundo máxima distância	80
Figura 43 – Distribuição de valores para o coeficiente de menor variação segundo MAD e segundo máxima distância	81
Figura 44 – Reconstruções da superfície 0 de 2LJ5 com seleção de coeficientes segundo MAD após filtragem por Daubechies	82
Figura 45 – Reconstruções da superfície 71 de 2LJ5 com seleção de coeficientes segundo MAD após filtragem por Daubechies	83
Figura 46 – RMSD de reconstruções com MAD e filtragem Daubechies de 2LJ5 com número variado de coeficientes da estrutura de referência	83
Figura 47 – Curvas de MAD a partir de valores originais e após filtragem por <i>wavelet</i> de Daubechies.	85
Figura 48 – RMSD de reconstruções de 2LJ5 com número variado de coeficientes da estrutura de referência com seleção de coeficientes segundo MAD e segundo influência no RMSD	86
Figura 49 – Reconstruções da superfície 0 de 2LJ5 com ordenação de coeficientes segundo influência no RMSD	86

Figura 50 – Reconstruções da superfície 71 de 2LJ5 com ordenação de coeficientes segundo influência no RMSD	87
Figura 51 – Espirais retangulares de sentido horário	94
Figura 52 – Superfície voxelizada de 2K5X_A e reconstruções em Zernike 3D	97
Figura 53 – Valores reconstruídos de 2K5X_A em voxelização binária por SASTBX . .	98
Figura 54 – Reconstruções de uma função de exemplo pelo programa de Nisius e Gohlke	100
Figura 55 – Valores de <i>voxel</i> para as coordenadas X de 2K5X_A calculados por SASTBX	101
Figura 56 – Valores de <i>voxel</i> para as coordenadas Y de 2K5X_A calculados por SASTBX	102
Figura 57 – Valores de <i>voxel</i> para as coordenadas Z de 2K5X_A calculados por SASTBX	103
Figura 58 – Reconstruções dos componentes coordenados de 2K5X_A pelo programa de Nisius e Gohlke	104
Figura 59 – Reconstrução de 2K5X_A na abordagem com ordenação axial	105
Figura 60 – Valores de <i>voxel</i> da coordenada X da superfície de 2K5X_A na abordagem de mapeamento na bola com reordenação axial.	106
Figura 61 – Valores de <i>voxel</i> da coordenada Y da superfície de 2K5X_A na abordagem com ordenação axial	107
Figura 62 – Valores de <i>voxel</i> da coordenada Z da superfície de 2K5X_A na abordagem com ordenação axial	108
Figura 63 – Função <i>X</i> da superfície de 2K5X_A visitada seguindo uma espiral e sua reconstrução	109
Figura 64 – Função <i>Y</i> da superfície de 2K5X_A visitada seguindo uma espiral e sua reconstrução	110
Figura 65 – Função <i>Z</i> da superfície de 2K5X_A visitada seguindo uma espiral e sua reconstrução	110
Figura 66 – Reconstrução de 2K5X_A pelo método de ordenação espiral	111

LISTA DE TABELAS

Tabela 1	–	Interpretação geométrica de curvaturas Gaussiana e Média em superfícies.	43
Tabela 2	–	Detalhes de estruturas selecionadas	46
Tabela 3	–	Métricas de distorção para os mapeamentos ConformalizedMCF, BLD e AHSP da superfície de 1R2M.	65
Tabela 4	–	Métricas de distorção para os mapeamentos ConformalizedMCF, BLD e AHSP da superfície de 2K5X_A.	65
Tabela 5	–	Detalhes de estruturas selecionadas	74
Tabela 6	–	Resultados de integração para duas definições de polinômios tridimensionais de Zernike.	96

LISTA DE ABREVIATURAS E SIGLAS

3D	Tridimensional
ARAP	As Rigid As Possible
BLD	Bijective and Low Distortion
CALD	Control of Area and Length Distortion
CGAL	Computational Geometry Algorithms
ESBTL	Easy Structural Biology Template Library
IRF	Iterative Residual Fitting
MAD	Mean Absolute Deviation
MCF	Mean Curvature Flow
MS	Molecular Surface
PDB	Protein Data Bank
RMSD	Root Mean Squared Deviation
SASTBX	Small Angle Scattering Toolbox

SUMÁRIO

1	INTRODUÇÃO	23
1.1	A importância da determinação de complexos entre proteínas	23
1.2	Representação de superfícies moleculares	26
1.3	Harmônicos esféricos	27
1.4	Polinômios tridimensionais de Zernike	30
1.5	Objetivos	32
2	REPRESENTAÇÃO DE SUPERFÍCIES MOLECULARES EM HARMÔ- NICOS ESFÉRICOS	33
2.1	Ambiente computacional	33
2.2	Ortonormalidade de harmônicos esféricos	34
2.3	Obtenção de malha triangular	34
2.3.1	Surfmap	34
2.3.2	ESBTL e CGAL	35
2.3.3	EDTSurf	35
2.4	Identificação e fechamento de túneis	35
2.5	Mapeamento de superfícies moleculares na esfera unitária	36
2.5.1	Surfmap	36
2.5.2	CALD	36
2.5.3	ConformalizedMCF	37
2.5.4	ARAP	37
2.5.5	BLD	38
2.5.6	AHSP	38
2.5.7	Métrica de distorção de área	38
2.6	Cálculo de coeficientes de harmônicos esféricos	39
2.7	Reconstrução e visualização de superfícies	41
2.8	Cálculo de curvaturas da superfície	42
2.9	Cálculo de energia potencial eletrostática	44
2.10	Decomposição espectral de superfícies de um <i>ensemble</i> de confor- mações	45
2.11	Conjunto de dados	46
3	RECONSTRUÇÃO DE SUPERFÍCIES MOLECULARES EM HARMÔ- NICOS ESFÉRICOS	47
3.1	Ortonormalidade de harmônicos esféricos	47
3.2	Reconstrução de superfícies em harmônicos esféricos	48

3.2.1	Surfmap e integração numérica	48
3.2.2	ESBTL, CGAL, CALD e IRF	51
3.2.3	ESBTL, CGAL, ConformalizedMCF e IRF	54
3.2.4	Triangulação pronta e IRF	55
3.2.5	EDTSurf, ConformalizedMCF e IRF	56
3.2.6	ARAP e IRF	58
3.2.7	EDTSurf, BLD e IRF	61
3.2.8	EDTSurf, AHSP e IRF	61
3.3	Distorções do mapeamento	62
3.4	Detecção e fechamento de túneis	66
3.5	Reconstrução e visualização de superfícies	69
3.6	Cálculo de curvaturas da superfície	74
3.7	Cálculo de energia potencial eletrostática	75
3.8	Decomposição espectral de superfícies de um <i>ensemble</i> de conformações	77
3.9	Conclusões	84
4	REPRESENTAÇÃO DE SUPERFÍCIES MOLECULARES EM ZER- NIKE 3D COM MAPEAMENTO NA BOLA UNITÁRIA	89
4.1	Avaliação de integração	89
4.2	Voxelização	90
4.3	Reconstrução de objetos volumétricos binários	91
4.4	Reconstrução de objetos volumétricos definidos por funções gerais	92
4.5	Mapeamento na bola unitária com reordenação axial	92
4.6	Mapeamento na bola unitária com reordenação espiral	93
5	RECONSTRUÇÃO DE SUPERFÍCIES MOLECULARES EM ZER- NIKE 3D	95
5.1	Avaliação de integração	95
5.2	Reconstrução de funções binárias	95
5.3	Reconstrução de funções gerais	99
5.4	Representação em Zernike 3D com reordenação axial	104
5.5	Representação em Zernike 3D com reordenação espiral	109
6	CONCLUSÃO	113
	REFERÊNCIAS	115

1 INTRODUÇÃO

1.1 A importância da determinação de complexos entre proteínas

O desenvolvimento dos computadores permitiu grandes avanços nas áreas de ciência e tecnologia. Com o suporte provido pelos computadores, grandes problemas científicos e de engenharia são resolvidos e novos são propostos. Uma área de destaque é a simulação de sistemas naturais, como evolução estelar, modelagem climática e, especialmente, modelagem de sistemas biológicos. Em um organismo vivo, a grande maioria dos processos celulares, incluindo catálise de enzimas, regulação de expressão gênica e resposta imunológica, depende da formação de complexos macromoleculares permanentes ou transientes envolvendo proteínas.¹ Proteínas são longas cadeias de resíduos de aminoácidos unidos por ligações peptídicas; cada ligação une um par de aminoácidos e elimina uma molécula de água - daí o termo *resíduo*.

As funções biológicas das proteínas dependem de suas estruturas tridimensionais e de interações físicas e diretas com outras moléculas,² através de um equilíbrio delicado de complicadas interações atômicas e de solvatação na interface. Ao contrário de outras macromoléculas biológicas, proteínas apresentam estruturas complexas e irregulares. Sua capacidade de apresentar superfícies estrutural e quimicamente diversas permite alta especificidade nas interações.^{3,4} O papel das proteínas nos organismos é tão crucial que complexos proteicos são descritos como máquinas moleculares,⁵ e problemas relacionados à formação de tais complexos estão associados a doenças graves, como anemia falciforme e alguns tipos de câncer.⁴ A determinação de como proteínas interagem permite compreender melhor uma variedade de processos celulares no nível molecular.

Informação estrutural sobre tais complexos, ou seja, as coordenadas tridimensionais dos núcleos de seus átomos constituintes, em resolução baixa a intermediária, pode ser obtida por meio de várias técnicas experimentais, como, por exemplo, ressonância magnética nuclear^{6,7} e mais recentemente por microscopia crio-eletrônica.⁸ Entretanto, quando se deseja estruturas em alta resolução, a cristalografia por raios X é a técnica mais estabelecida.⁹ Há, contudo, casos em que se conhece a estrutura das unidades moleculares mas não a do complexo,¹⁰ e casos em que um cristal pode ser obtido para um complexo mas não em uma conformação biologicamente relevante. Ademais, crê-se que a maioria das interações funcionais entre proteínas são transientes, formando complexos que não são suficientemente estáveis para cristalização.¹¹ A dificuldade de determinação experimental de estruturas de alguns complexos combinada com o crescente número de estruturas tridimensionais de proteínas individuais estimulou o emprego de modelagem computacional para dedução da estrutura de complexos.

A modelagem de interações entre proteínas pode ser abordada sob ao menos duas diferentes perspectivas. A primeira delas, simulação da associação (ou *binding*), visa determinar o modo como tais macromoléculas interagem em termos cinéticos e termodinâmicos e auxilia a compreensão dos mecanismos de formação de complexos e do fluxo de informação e de materiais em sistemas biológicos.^{11,12} A correta caracterização das trajetórias de aproximação das moléculas envolvidas exige simulações que considerem interações entre átomos individuais das moléculas e do solvente (*all-atom protein interactions*) e resultem na cinética e termodinâmica do sistema. Com complexidade comparável à da simulação do enovelamento de proteínas,^{10,13} i.e. de como proteínas adquirem seu formato tridimensional a partir de uma mera sequência de resíduos de aminoácidos, a simulação de associação de proteínas foi por muito tempo considerada intratável, mesmo no caso de interações entre proteína e pequenos ligantes.¹⁴ Avanços em arquiteturas de computadores e em algoritmos têm permitido simulações de descrição atômica completa referentes a escalas de tempo significativas,^{15,16} no entanto acessível apenas através de capacidade computacional de altíssimo nível.¹⁷ Ainda assim, vários cálculos de trajetórias são tipicamente exigidos para amostragem estatisticamente significativa do espaço conformacional das moléculas, impondo marcante demanda por recursos computacionais.¹⁸

Alternativamente, a simulação do atracamento de proteínas, também conhecida como *docking* macromolecular, pretende especificar a configuração final de associação intermolecular sem considerar o processo prévio de combinação. É, por isso, computacionalmente mais eficiente e permite busca por uma diversidade maior de candidatos a complexos. O trabalho pioneiro de [Levinthal et al.](#)¹⁹ inaugurou o uso de programas de computador para determinação de complexos de proteínas, através do cálculo do arranjo de moléculas de hemoglobina falciforme a partir de análise teórica de possíveis modos de agregação de hemoglobina em fibras regulares. Foi sucedido pela determinação dos possíveis modos de interação entre tripsina, uma proteína do sistema digestivo, e o inibidor de tripsina pancreática bovina,¹⁰ estabelecendo, com isso, os fundamentos de *docking* de proteínas. Desde então, vários programas de computador têm sido desenvolvidos com o objetivo de calcular a estrutura tridimensional de um complexo proteico a partir das estruturas de seus componentes individuais.²⁰

Três componentes básicos são tipicamente observados numa metodologia de *docking* de proteínas: representação das moléculas, busca por candidatos, e avaliação das interações. A *representação* de proteínas adotada pelo programa deve reproduzir adequadamente as propriedades das moléculas relevantes para os cálculos de arranjo. Uma vez apropriadamente representadas, as moléculas são então sujeitas a repetições das etapas de *busca* por candidatos a complexos, através de orientação e posicionamento das moléculas envolvidas, e de *avaliação* da interação. Na etapa de busca, a orientação das moléculas é decidida pelo algoritmo, e então operações de rotação e de translação são aplicadas sobre as estruturas tridimensionais para gerar as possíveis interações entre as moléculas e simular o encaixe.

A posição relativa das moléculas é descrita por 6 variáveis, correspondentes a 3 graus de liberdade de rotação e 3 graus de liberdade de translação. Posto que a configuração final do complexo é desconhecida, um vasto espaço de busca precisa ser examinado para a determinação das orientações mais favoráveis. Na segunda etapa, funções potenciais avaliam a intensidade da interação e guiam a etapa seguinte de orientação.

Proteínas, contudo, não são moléculas totalmente rígidas. Macromoléculas em solução manifestam vibrações causadas por flutuações termodinâmicas intrínsecas,²¹ em variadas amplitudes, coletivamente denominadas de flexibilidade estrutural. Ainda que técnicas experimentais como cristalografia de raios-X identifiquem uma conformação (configuração dos átomos), macromoléculas podem assumir elevado número de conformações, chamados de subestados conformacionais,²² sendo o de maior frequência tipicamente observado experimentalmente. A diversa população de uma mesma proteína é agrupada em uma coleção chamada de *ensemble* conformacional.^{23,24} Além de vibrações intrínsecas, também pode haver margens consideráveis de variação conformacional durante a formação do complexo proteico, fazendo com que a dificuldade imposta pela flexibilidade estrutural das moléculas seja um dos principais desafios da predição da estrutura final do complexo,^{25,26} ao lado da diferenciação entre estruturas calculadas similares ao complexo natural²⁷ e da identificação de resíduos mais influentes na estabilização da interface entre proteínas.²⁸

Diversos programas de computador têm sido propostos para simular o atracamento de proteínas, com emprego de técnicas variadas para cada um dos três componentes básicos e com grau de sucesso igualmente diverso.^{20,29,30} De modo geral, as metodologias podem ser reunidas em duas categorias de acordo com a estratégia de tratamento de flexibilidade estrutural. No *docking* rígido, é assumido que as moléculas possuem estrutura estática e qualquer mudança conformacional é ignorada. Já as técnicas de *docking* flexível surgiram com a compreensão da importância das variações conformacionais. Comum a muitas abordagens de *docking*, flexibilidade pode ser tratada de modo explícito, na forma de uma etapa final de refinamento por simulações de dinâmica molecular (DM). O grau de flexibilidade comportado, e portanto também o custo computacional associado, varia desde discretos movimentos da cadeia lateral até movimentos completos da cadeia principal, possivelmente resultando em alterações conformacionais acentuadas. Por outro lado, flexibilidade pode ser tratada de modo implícito, como é o caso de *soft docking*, que altera o *docking* rígido com a permissão de alguma sobreposição das moléculas; e do chamado *ensemble docking*, uma repetição de *docking* rígido com várias combinações de estruturas de um *ensemble* conformacional. O acréscimo esperado do custo computacional em comparação com *docking* rígido tradicional justifica-se pelo acesso a um panorama energético mais amplo^{21,31} e consequente probabilidade de encontro da configuração de complexo semelhante à natural. Este é o método adotado por CamDock.³²

1.2 Representação de superfícies moleculares

Uma vez que a configuração final do complexo proteico é desconhecida, o sucesso do atracamento requer múltiplas reorientações das moléculas manipuladas. A demanda por recursos computacionais, tanto de armazenamento das estruturas moleculares quanto de seu processamento, deve, portanto, ser reduzida para o *docking* ser mais atraente. Isto é ainda mais crítico em *ensemble docking*, em que múltiplas conformações são examinadas. Por outro lado, a precisão das manipulações precisa ser elevada para os resultados atingidos serem corretos. Estruturas de moléculas biológicas obtidas experimentalmente são descritas em PDB, o formato do *Protein Data Bank* (banco de dados de proteínas),³³ um repositório de estruturas livremente acessível para a comunidade científica mundial. As informações estruturais são organizadas em arquivos de formato texto e indicam os resíduos e átomos de uma proteína e suas respectivas coordenadas no espaço tridimensional, além de dados sobre o método específico de determinação de sua estrutura.³⁴ Modelos atômicos como o PDB são vantajosos no exame de interações internas à molécula; porém, não são convenientes para a reprodução de características da superfície molecular, da qual dependem as interações entre proteínas.¹⁰ Uma descrição de superfície molecular se faz, portanto, necessária.

Superfície molecular é compreendida como um modelo representando um envólucro que engloba todos os átomos da molécula. Lee e Richards³⁵ propuseram um método em que átomos da superfície podem ser distinguidos de átomos internos e a superfície total da molécula pode ser descrita como a soma das superfícies atômicas. Desde então, três tipos de superfície são tipicamente utilizadas:^{36,37} a superfície de van der Waals, em que cada átomo é representado por uma esfera rígida com raio determinado pelo raio do átomo; a superfície acessível ao solvente (*solvent-accessible surface*, ou SAS), formada pelo deslizamento de uma esfera de 1,4 Å de raio, representando uma molécula de água, pela camada externa dos átomos da molécula de interesse e considerando o envólucro formado pelo caminho realizado pelo ponto central da esfera; e superfície de Connolly, ou apenas superfície molecular (*molecular surface*, ou ainda *solvent-excluded surface*), obtida de forma semelhante à da superfície acessível ao solvente, porém considerando o envólucro formado pelo contato entre a esfera de sondagem e os átomos da molécula. O modelo de superfície molecular é uma descrição interessante por permitir uma representação conveniente da nuvem de elétrons em volta dos átomos. Tipicamente, tais superfícies são descritas de forma discreta como uma coleção de pontos no espaço tridimensional. O elevado detalhamento de descrições pontuais de superfícies moleculares amplifica, contudo, a demanda por recursos computacionais.

A fim de abrandar a complexidade dos cálculos de simulação de atracamento, Wodak e Janin¹⁰ lançaram mão de “suposições simplificadoras cruciais” [p. 324, tradução nossa], a começar pela omissão de flexibilidade estrutural nos estágios iniciais de *docking*. Um modelo simplificado representou as proteínas como uma coleção de esferas centra-

Figura 1 – Projeção de superfícies em harmônicos esféricos com diferentes ordens de aproximação L_{max} .

Fonte: GREEN.⁴⁵

das em cada resíduo. Um mapa topográfico da superfície foi utilizado para facilmente localizar características da superfície em baixo detalhamento. O influente algoritmo de reconhecimento geométrico proposto por [Katchalski-Katzir et al.](#),³⁸ por sua vez, embora capaz de identificar complementaridade de superfícies moleculares e aproveitar correlações e transformadas de Fourier para diminuir os passos de computação, ainda necessita que cada superfície de um *ensemble* conformacional seja descrita individualmente. Portanto, uma descrição mais adequada de superfície molecular é necessária.

1.3 Harmônicos esféricos

Harmônicos esféricos são um conjunto de funções que têm permitido diversas aplicações, como na visualização de imagens médicas,^{39–41} na representação de superfícies gerais,⁴² na análise de densidades de carga,⁴³ no cálculo de distribuição de densidade de ligantes em superfícies moleculares⁴⁴ e em iluminação em jogos eletrônicos.⁴⁵ Também foram usados na representação de superfícies moleculares^{46,47} e em *docking* de proteínas,^{32,48} com aproveitamento de suas operações eficientes de rotação.^{32,49} Como exemplificado pela figura 1, o uso de harmônicos esféricos permite refinamento sucessivo da representação de superfícies.⁴⁵

Na esfera unitária, os valores θ e ϕ representam, respectivamente, o ângulo polar e o ângulo azimutal no sistema de coordenadas esféricas, com $\theta \in [0; \pi]$ e $\phi \in [0; 2\pi]$. Os harmônicos esféricos $Y_l^m(\theta, \phi)$ de grau l e ordem m são a parte angular da solução da equação de Laplace,⁵⁰ $\nabla^2 f = 0$, e constituem um conjunto ortonormal de funções definidas na superfície esférica. Sua forma analítica é definida pela equação 1.1 para $l \in \mathbb{Z}^+$,

$m \in \mathbb{Z}$, $-l \leq m \leq l$; m é um índice, não um expoente.⁵¹

$$Y_l^m(\theta, \phi) = \sqrt{\frac{2l+1}{4\pi} \frac{(l-m)!}{(l+m)!}} P_l^m(\cos \theta) e^{im\phi} \quad (1.1)$$

P_l^m são os polinômios associados de Legendre, definidos pela equação 1.2, e P_n são os polinômios de Legendre, dados pela equação 1.3.

$$P_l^m(x) = (-1)^m (1-x^2)^{\frac{m}{2}} \frac{d^m}{dx^m} P_l(x) \quad (1.2)$$

$$P_n(x) = \frac{1}{2^n n!} \left(\frac{d}{dx} \right)^n (x^2 - 1)^n \quad (1.3)$$

$$\int_0^{2\pi} \int_0^\pi Y_l^m(\theta, \phi) \overline{Y_{l'}^{m'}(\theta, \phi)} \sin \theta d\theta d\phi = \delta_{ll'} \delta_{mm'} \quad (1.4)$$

$$\delta_{aa'} = \begin{cases} 1, & a = a' \\ 0, & a \neq a' \end{cases} \quad (1.5)$$

A relação de ortonormalidade dos harmônicos esféricos é dada pela equação 1.4, onde δ é o delta de Kronecker da equação 1.5. Por serem ortonormais, os harmônicos esféricos formam a base de um espaço de funções e podem ser usados para representar uma função qualquer $f(\theta, \phi)$, também definida na esfera, conforme equação 1.6.⁵² Os coeficientes λ_l^m indicam o peso de cada harmônico esférico na representação da função f e são calculados pela projeção de f no harmônico esférico correspondente, conforme equação 1.7.

$$f(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^l \lambda_l^m Y_l^m(\theta, \phi) \quad (1.6)$$

$$\lambda_l^m = \langle f, Y_l^m \rangle = \int_0^{2\pi} \int_0^\pi f(\theta, \phi) \overline{Y_l^m(\theta, \phi)} \sin \theta d\theta d\phi \quad (1.7)$$

Em cálculos numéricos, limita-se o parâmetro l a uma banda máxima L_{max} , também chamada de *ordem da expansão*,⁴³ para ser possível *aproximar* a função f pela reconstrução f_{SH} , de acordo com a fórmula 1.8, utilizando um computador eletrônico. O valor de L_{max} depende do nível de detalhamento que se deseja obter na reconstrução de f , uma vez que o erro na expansão diminui com o aumento de l .

$$f(\theta, \phi) \approx f_{SH}(\theta, \phi) = \sum_{l=0}^{L_{max}} \sum_{m=-l}^l \lambda_l^m Y_l^m(\theta, \phi) \quad (1.8)$$

Y_l^m são funções de imagem complexa. A representação de funções reais pode ser realizada também com harmônicos esféricos reais, que são obtidos a partir de combinações lineares dos harmônicos esféricos complexos, resultando na equação 1.9⁴⁵ para $K_l^m = \sqrt{\frac{2l+1}{4\pi} \frac{(l-m)!}{(l+m)!}}$.

$$y_l^m(\theta, \phi) = \begin{cases} \sqrt{2}K_l^m \sin(-m\phi)P_l^{-m}(\cos \theta), & m < 0 \\ K_l^0 P_l^0(\cos \theta), & m = 0 \\ \sqrt{2}K_l^m \cos(m\phi)P_l^m(\cos \theta), & m > 0 \end{cases} \quad (1.9)$$

Ritchie e Kemp⁴⁸ e Montalvao et al.³² utilizaram uma descrição radial de superfície para representar moléculas em harmônicos esféricos, em que a distância de seu centro geométrico até a superfície foi projetada neste espaço de funções. A representação em harmônicos esféricos exige que a superfície molecular seja do tipo *star-like*, ou seja, cada raio partindo do seu centro geométrico deve cruzar a superfície apenas uma vez.⁴⁶ Contudo, superfícies moleculares têm relevo acidentado, com saliências, depressões e túneis, devido ao contorno e posicionamento dos átomos, resultando em distorções na representação radial em harmônicos esféricos em regiões que não obedeçam o critério de *star-likeness*. A representação paramétrica em harmônicos esféricos,^{41,46} no entanto, evita a restrição de unicidade de cruzamentos de raios com a superfície através da combinação de duas técnicas. A primeira delas é decomposição da função $\mathbf{X}$ que define seu formato em três funções,⁴⁶ $X(\theta, \phi)$, $Y(\theta, \phi)$ e $Z(\theta, \phi)$, cada uma delas contendo as coordenadas da superfície referente a um eixo coordenado, resultando na equação 1.10.

$$\mathbf{X}(\theta, \phi) = \begin{bmatrix} X(\theta, \phi) \\ Y(\theta, \phi) \\ Z(\theta, \phi) \end{bmatrix} \quad (1.10)$$

Superfícies representadas de forma paramétrica são então descritas por três conjuntos de coeficientes de harmônicos esféricos, $\Lambda_x = \{\lambda_{x,l}^m\}$, $\Lambda_y = \{\lambda_{y,l}^m\}$ e $\Lambda_z = \{\lambda_{z,l}^m\}$. Cada função coordenada é reconstruída individualmente pela equação 1.8 com harmônicos esféricos reais.

Porém, as funções X , Y e Z podem não ter valor único no domínio da esfera unitária. A segunda técnica necessária para a representação paramétrica é a associação dos valores dessas funções com posições na esfera. Um mapeamento $\Psi : \mathbf{R}^3 \rightarrow \mathbf{R}^3$ é um operador que deforma a superfície de interesse, de coordenadas (x_m, y_m, z_m) , em uma esfera de coordenadas (x_s, y_s, z_s) , associando, com isso, a cada valor de (θ, ϕ) um único valor de (x_m, y_m, z_m) . Esta operação exige que a superfície original seja topologicamente equivalente a uma esfera. No entanto, eventuais túneis presentes na superfície impedem sua transformação num formato esférico. O termo matemático *genus* indica a “quantidade de curvas simples fechadas disjuntas duas a duas que existem em uma superfície M que,

coletivamente, não a desconectam”⁵³ e captura a ideia de túneis em uma superfície; seu valor equivale ao número de alças adicionadas a uma esfera para resultar na superfície de interesse.⁵³ Apenas superfícies de *genus* 0 são representáveis nesta abordagem.

A representação paramétrica em harmônicos esféricos é capaz de descrever detalhes com resolução ajustável e respectivo volume de dados também ajustável. Foi utilizada para visualização de superfícies^{46,54} e se mostrou apropriada para uso em atracamento de proteínas⁵⁵ devido a menores requisitos de memória e por tirar proveito de maior eficiência de rotação de harmônicos esféricos em comparação a pontos no espaço cartesiano.^{45,49,56} Cabem, contudo, melhorias em sua qualidade de representação e em sua aplicabilidade a superfícies de *genus* maior que zero.

1.4 Polinômios tridimensionais de Zernike

Polinômios de Zernike são também um conjunto de funções ortogonais, porém definidas no círculo unitário. Foram inicialmente propostos como uma solução para a determinação de aberrações em um instrumento óptico⁵⁷ por permitir isolar e quantificar cada aberração separadamente. Além da área de óptica, foram aplicados com sucesso em análise de imagens,⁵⁸ em especial em imagens médicas.⁵⁹ São formulados como o produto de um componente radial e de um componente angular. Os componentes angulares formam um conjunto ortogonal de funções e os componentes radiais são tais que seu produto pelos angulares seja, também, um conjunto ortogonal. A ortogonalidade de polinômios de Zernike permite que sejam usados, assim como os harmônicos esféricos, para reconstruir funções arbitrárias como uma combinação linear de polinômios de Zernike, em que cada polinômio é multiplicado por um coeficiente, também chamado de *momento* de Zernike. Cada momento representa uma projeção da função de interesse no conjunto de funções dado pelos polinômios de Zernike.

Embora originalmente bidimensionais, polinômios de Zernike podem ser estendidos para dimensões arbitrárias. Canterakis⁶⁰ formulou-os para três dimensões. Segundo sua definição, polinômios tridimensionais de Zernike Z_{nl}^m têm como componente angular os harmônicos esféricos complexos Y_l^m da seção 1.3 multiplicados por uma função radial R_{nl} conforme a equação 1.11, para $\rho \in [0, 1]$, $\theta \in [0, \pi]$ e $\phi \in [0, 2\pi]$. Os índices inteiros n , l e m respeitam $n \geq 0$, $0 \leq l \leq n$, $-l \leq m \leq l$ e $n - l$ deve ser par.

$$Z_{nl}^m(\rho, \theta, \phi) = R_{nl}(\rho)Y_l^m(\theta, \phi) \quad (1.11)$$

O componente R_{nl} é um fator real associado ao raio da posição de interesse e obtido de modo a tornar os polinômios de Zernike ortogonais na esfera unitária. A equação 1.12 define polinômios 3D (tridimensionais) de Zernike em coordenadas cartesianas, para

$\mathbf{x} \in \mathbb{R}^3$, com $\mathbf{x} = r\boldsymbol{\xi} = r(\sin \theta \cos \phi, \sin \theta \sin \phi, \cos \theta)$, $|\mathbf{x}| = r$ e $|\boldsymbol{\xi}| = 1$.

$$Z_{nl}^m(\mathbf{x}) = R_{nl}(r)Y_l^m(\boldsymbol{\xi}) \quad (1.12)$$

Já os momentos Ω_{nl}^m de Zernike 3D para uma função arbitrária $f(\mathbf{x})$ são calculados pela equação 1.13. Tais valores são utilizados na reconstrução da função $f(\mathbf{x})$ conforme a equação 1.14. Uma aproximação desta função, $f_{ZK}(\mathbf{x})$, pode ser obtida pela limitação da *ordem de expansão* n a um parâmetro N_{max} , conforme equação 1.15. Tanto o cálculo de momentos quanto a reconstrução da função são semelhantes ao caso de harmônicos esféricos, devido à ortogonalidade dos polinômios de Zernike, equação 1.16, e ortogonalidade das funções radiais, equação 1.17.

$$\Omega_{nl}^m = \frac{3}{4\pi} \int_{|\mathbf{x}| \leq 1} f(\mathbf{x}) \overline{Z_{nl}^m(\mathbf{x})} d\mathbf{x} \quad (1.13)$$

$$f(\mathbf{x}) = \sum_{n=0}^{\infty} \sum_{l=0}^n \sum_{m=-l}^l \Omega_{nl}^m Z_{nl}^m(\mathbf{x}) \quad (1.14)$$

$$f(\mathbf{x}) \approx f_{ZK}(\mathbf{x}) = \sum_{n=0}^{N_{max}} \sum_{l=0}^n \sum_{m=-l}^l \Omega_{nl}^m Z_{nl}^m(\mathbf{x}) \quad (1.15)$$

$$\frac{3}{4\pi} \int_{|\mathbf{x}| \leq 1} Z_{nl}^m(\mathbf{x}) \cdot \overline{Z_{n'l'}^{m'}(\mathbf{x})} d\mathbf{x} = \delta_{nn'} \delta_{ll'} \delta_{mm'} \quad (1.16)$$

$$\int_0^1 R_{nl}(r) R_{n'l}(r) r^2 dr = \delta_{nn'} \quad (1.17)$$

Canterakis⁶⁰ mostrou que o componente radial dos polinômios 3D de Zernike pode ser descrito pela equação 1.18, cujos coeficientes q_{kl}^v são dados pela equação 1.19.

$$R_{nl}(r) = r^l \sum_{v=0}^k q_{kl}^v r^{2v}, \quad k = \frac{n-l}{2} \quad (1.18)$$

$$q_{kl}^v = \frac{(-1)^k}{2^{2k}} \sqrt{\frac{2l+4k+3}{3}} \binom{2k}{k} (-1)^v \frac{\binom{k}{v} \binom{2(k+l+v)+1}{2k}}{\binom{k+l+v}{k}} \quad (1.19)$$

Como polinômios de Zernike 3D têm como domínio a bola unitária, pode ser necessário multiplicar a função ou o objeto a reconstruir por um fator de escala de modo a contê-los neste espaço.

A presença de um componente radial nos polinômios de Zernike torna possível identificar univocamente cada ponto na bola através da tupla (ρ, θ, ϕ) (o que inclui o interior da esfera). Esta propriedade leva a duas consequências importantes. A primeira é

a eliminação da necessidade de mapeamento esférico vista na seção 1.3 para representação paramétrica em harmônicos esféricos, uma vez que a localização dos pontos não se limita mais à posição angular isolada (θ, ϕ) que os identifica apenas na esfera. A segunda consequência é a capacidade de representar túneis na superfície molecular. Reconstrução em Zernike 3D não se limita, portanto, a superfícies de *genus* 0 e pode ser calculada para superfícies com qualquer topologia. Ambas características são muito apreciadas, dadas a complexidade de mapeamentos esféricos e a variedade de acidentes no relevo de superfícies moleculares.

Esta representação foi utilizada para descrição de superfícies moleculares⁶¹ e *docking* de proteínas⁶²; no entanto, seu uso foi limitado a funções binárias, o que limita as características da superfície que podem ser representadas. Cabe, por conseguinte, investigação de seu uso em reconstrução de funções gerais.

1.5 Objetivos

O objetivo principal deste projeto é o estabelecimento de métodos computacionais para a representação de superfícies moleculares em alta qualidade adequadas para aplicação em simulação de formação de complexos entre proteínas com tratamento de flexibilidade implícito através de *ensemble* conformacional. A fim de alcançá-lo, os seguintes objetivos específicos são propostos:

- modernizar a representação de superfícies moleculares paramétricas em harmônicos esféricos⁴⁶ com a adoção de algoritmos mais recentes;
- generalizar a representação de superfícies moleculares paramétricas em harmônicos esféricos para superfícies de qualquer *genus*;
- estender a representação de imagens volumétricas em polinômios tridimensionais de Zernike^{60,63} com o acréscimo de bases de funções e consequente elevação da fidelidade na representação;
- explorar similaridade estrutural de um *ensemble* conformacional para reduzir demanda por memória e processamento;

Este texto está organizado do seguinte modo: o capítulo 2 apresenta a metodologia de análise da representação paramétrica em harmônicos esféricos. Seus resultados são apresentados e discutidos no capítulo 3. A representação em polinômios tridimensionais de Zernike é estudada no capítulo 4, cujos resultados são discutidos no capítulo 5. Por fim, o capítulo 6 expõe as conclusões.

2 REPRESENTAÇÃO DE SUPERFÍCIES MOLECULARES EM HARMÔNICOS ESFÉRICOS

A representação de superfícies proteicas em harmônicos esféricos de forma paramétrica decompõe o formato tridimensional em três funções relativas aos eixos coordenados. Simplificadamente, dada a descrição atômica de uma proteína, é necessário calcular uma malha triangular que defina sua superfície molecular, que deve ser posteriormente deformada num formato esférico para então ser projetada na base dos harmônicos esféricos. Finalmente, a reconstrução a partir dos coeficientes obtidos no passo anterior estabelece uma aproximação da superfície original.

Este capítulo apresenta a metodologia de investigação para obtenção de resultados de alta qualidade com esta representação. Diversos algoritmos e implementações foram testados; neste capítulo, são agrupados por categoria. A seção 2.1 inicia a apresentação com descrição do ambiente computacional utilizado. A introdução dos métodos para cálculo de superfície molecular e malha triangular é feita na seção 2.3. Em seguida, a seção 2.4 indica o procedimento para fechamento de túneis, sucedida pela descrição dos algoritmos de mapeamento esférico na seção 2.5. O cálculo de coeficientes de harmônicos esféricos é indicado na seção 2.6, e a seção 2.7 finaliza o estágio de representação com detalhes sobre a reconstrução de superfícies. As curvaturas que definem o relevo das superfícies são obtidas pelo procedimento da seção 2.8, ao passo que o método de obtenção da energia potencial eletrostática é apresentado na seção 2.9. A seção 2.10 descreve a decomposição espectral para economia de coeficientes em um *ensemble* de conformações. O capítulo se encerra com a apresentação do conjunto de dados utilizado na seção 2.11.

2.1 Ambiente computacional

Um dos objetivos deste trabalho é que as ferramentas computacionais geradas sejam utilizadas por outros grupos de pesquisa. Para tanto, questões de portabilidade foram constantemente consideradas em seu desenvolvimento, como forte adesão aos padrões e limitação do uso de funcionalidades sabidamente disponíveis em apenas alguns sistemas. Segundo Mooney,⁶⁴ “uma aplicação é portátil em uma classe de ambientes se o esforço requerido para transportá-la e adaptá-la para um novo ambiente na classe é menor do que o esforço para desenvolvê-la novamente” [p. 1, tradução nossa]. Neste sentido, a ampla utilização da ferramenta é favorecida com portabilidade do código-fonte. Para tanto, foi escolhida a linguagem de programação C, reconhecida pela sua disponibilidade na maioria dos sistemas computacionais, pela razoável gama de funcionalidades de sua biblioteca padrão, e pelo alto desempenho das aplicações geradas. Além disso, C é uma linguagem muito popular entre os desenvolvedores de aplicações, ocupando a primeira ou segunda posições na avaliação TIOBE desde ao menos o ano de 2002.⁶⁵ Os programas desenvolvidos

seguiram a versão C99, estabelecida em 1999, e a maioria dos recursos utilizados, tanto de funcionalidades da linguagem em si quanto de funcionalidades da biblioteca padrão, são da versão C90, de 1990. Alguns trechos pontuais foram escritos em C++ para permitir a interação com bibliotecas escritas nesta linguagem e seguem a versão C++11, do ano de 2011. Pelo mesmo motivo, outros trechos foram escritos em linguagem Python, versão 3, de 2008. Cabe ressaltar que tais preocupações colaboram para a portabilidade; contudo, não há uma garantia de que o sistema de software final executará em qualquer computador, uma vez que há uma ampla variedade de combinações de hardware e de software e o esforço para adaptar a ferramenta para um sistema computacional específico pode ser elevado.

O ambiente de desenvolvimento consistiu em computador com processador de quatro núcleos Intel Core i7-4700 e 32GB de memória RAM executando o sistema operacional GNU/Linux, com compilador GCC 4.8.4⁶⁶ e sistema de automação de compilação CMAKE 2.8.12.2.⁶⁷ Todas as bibliotecas utilizadas pelo programa disponibilizam seu código-fonte em licenças permissivas.

2.2 Ortonormalidade de harmônicos esféricos

Ao se usar funções matemáticas para resolver problemas, é importante garantir que sua implementação tenha as mesmas propriedades que os objetos matemáticos a que se referem, dadas as limitações de um computador eletrônico finito. Para tanto, foi verificada a relação de ortonormalidade de harmônicos esféricos, expressa pela equação 2.1.⁵¹

$$\int_0^{2\pi} \int_0^\pi y_l^m(\theta, \phi) y_{l'}^{m'}(\theta, \phi) \sin \theta d\theta d\phi = \delta_{ll'} \delta_{mm'} \quad (2.1)$$

2.3 Obtenção de malha triangular

2.3.1 Surfmap

Inicialmente, obteve-se a superfície molecular triangulada a partir do algoritmo implementado no programa Surfmap.^{68,69} Projetado para analisar estruturas de proteínas e de DNA, este programa realiza o cálculo de uma malha triangular, representando superfície molecular, e o cálculo de mapeamento desta superfície na esfera. A partir das coordenadas atômicas da proteína, obtidas de um arquivo PDB, uma grade cúbica é estabelecida e uma esfera com raio de 1,4 Å, representando uma molécula de água, é deslocada pela grade seguindo uma camada que considera os valores padrão de raios atômicos para formar um contorno tridimensional que, finalmente, gera uma superfície fechada e triangulada. Surfmap também calcula o mapeamento esférico, conforme seção 2.5.

2.3.2 ESBTL e CGAL

Em comparação a Surfinap, foi adotada a combinação de bibliotecas ESBTL (do inglês *Easy Structural Biology Template Library*, ou biblioteca fácil de formatos para biologia estrutural)⁷⁰ e CGAL (do inglês *Computational Geometry Algorithms Library*, ou biblioteca de algoritmos de geometria computacional).⁷¹ O objetivo do projeto CGAL⁷¹ é “fornecer acesso a algoritmos geométricos eficientes e confiáveis, na forma de uma biblioteca C++” [p. 1, tradução nossa]. O uso da biblioteca por diversos projetos de diversas áreas e o rigor da implementação de seus algoritmos foram um bom indicativo de sua qualidade. As coordenadas atômicas de um arquivo PDB foram obtidas com o auxílio da biblioteca ESBTL e encaminhadas para CGAL. Esta cria uma superfície do tipo envólucro, composta de cascas resultantes da união de esferas. Cada esfera de entrada é centrada em um átomo da molécula e seu raio é dado pelos valores padrão de raio atômico⁷² somados de 1,4 Å. O passo seguinte cria uma malha triangular a partir da superfície e pode ser reaplicado para se obter triangulações finas da superfície.

2.3.3 EDTSurf

Malhas triangulares de superfícies moleculares foram também geradas pelo programa EDTSurf.⁷³ Este programa identifica a distância mínima de pontos na molécula até a superfície para diferenciar pontos internos de externos e, com o auxílio de uma grade, aplica o algoritmo *marching cubes*⁷⁴ para converter um conjunto de pontos em uma malha conectada. É possível gerar os três tipos de superfície mencionados no capítulo 1, a saber superfície de van der Waals, superfície acessível ao solvente e superfície molecular (Connolly).

2.4 Identificação e fechamento de túneis

Superfícies moleculares têm relevo acidentado, devido ao contorno dos átomos, e exibem protruções, cavidades e túneis devido ao posicionamento dos átomos no espaço tridimensional. Ao se representar superfícies em harmônicos esféricos, é preciso considerar a presença de túneis. Para o mapeamento ser bem sucedido, é necessário que a superfície seja de *genus*⁵³ zero, isto é, topologicamente equivalente a uma esfera; intuitivamente, a molécula não deve conter túneis. O software ReebHanTun⁷⁵ foi utilizado para ajustar superfícies e permitir mapeamento esférico. Este programa colapsa o contorno da superfície em menos pontos para criar um grafo, a partir do qual as voltas de um *loop* em torno do outro são usadas para identificar *loops* de alça e *loops* de túnel. Após remoção dos vértices pertencentes a *loops* de túnel da malha da superfície, a rotina de fechamento de buracos de CGAL⁷⁶ foi usada para fechar a malha através de introdução de novos vértices e arestas na região previamente esvaziada. Uma vez que todos os túneis tenham sido assim eliminados, a malha está pronta para ser processada pelo estágio de mapeamento esférico.

2.5 Mapeamento de superfícies moleculares na esfera unitária

A representação de superfícies moleculares em harmônicos esféricos exige que a superfície seja do tipo *star-like*, ou seja, cada raio partindo do seu centro geométrico deve cruzar a superfície apenas uma vez.⁴⁶ A parametrização segue o domínio dos harmônicos esféricos, que é, naturalmente, a esfera unitária. Qualquer função $f(\mathbf{x}) : \mathbb{R}^2 \rightarrow \mathbb{R}$ a ser projetada deve apresentar unicidade de valores em todo este domínio, isto é, $\forall \mathbf{x} \in \mathbb{R}^2, f(\mathbf{x}_1) \neq f(\mathbf{x}_2) \implies \mathbf{x}_1 \neq \mathbf{x}_2$. Faz-se necessária a transformação da descrição de superfície molecular em uma descrição topologicamente equivalente, porém esférica, através de uma operação de mapeamento.

2.5.1 Surfmap

Na etapa final de Surfmap, a partir da superfície triangulada calculada, cada vértice é associado a uma posição na esfera unitária especificada pelos dois parâmetros de coordenadas esféricas θ e ϕ . Os polos norte e sul são escolhidos como os dois vértices mais distantes entre si. Novos valores para a coordenada z de cada vértice são calculados pela média da coordenada z dos vértices vizinhos. O valor do ângulo azimutal é calculado como a média dos valores de ϕ dos vértices vizinhos, iniciando no polo norte e percorrendo um meridiano de fronteira. O método de Mercator para mapeamento cartográfico é usado para calcular o ângulo polar, θ , a partir da coordenada z , como a solução para a equação 2.2, onde C é uma constante positiva.

$$z = C \log \left(\frac{\cos \theta + 1}{\sin \theta} \right) \quad (2.2)$$

2.5.2 CALD

Surgido na área de imagens médicas, o método CALD (do inglês *Control of Area and Length Distortions*, ou controle de distorções de área e de comprimento)⁷⁷ foi projetado para gerar um mapeamento com controle de distorções de área e de comprimento. Um mapeamento inicial é realizado; em seguida, um estágio iterativo busca melhorar gradualmente a qualidade do mapeamento para obter uma melhor parametrização, através do reposicionamento dos vértices da malha triangular, em um procedimento chamado de suavização da malha.

Sua etapa de mapeamento inicial é uma extensão do trabalho de Brechbühler, Gerig e Kübler.⁴¹ Uma malha triangular que descreve uma superfície é submetida ao algoritmo e os polos norte e sul são escolhidos como os vértices mais distantes entre si. Em seguida, a colatitude θ é calculada para todos os vértices através da solução de um sistema de equações lineares $\mathbf{A}\boldsymbol{\theta} = \mathbf{b}$, sendo que $\theta_{\text{norte}} = 0$ e $\theta_{\text{sul}} = \pi$. Este sistema de equações é formado considerando que a colatitude de um vértice deve ser igual à média ponderada

dos vértices vizinhos, em que os pesos são inversamente proporcionais à distância de cada vizinho. A latitude ϕ é obtida de forma parecida, porém considerando a descontinuidade equivalente à linha internacional de data, em que ϕ tem seu valor diminuído em 2π . A linha de data é escolhida como a descida mais íngreme entre os polos, isto é, o caminho com maiores diferenças do ângulo θ entre vértices consecutivos. Para melhor distribuir a colatitude, a longitude é recalculada para dois mapeamentos, rotacionados um em torno do eixo X e o outro em torno de Y, e o mapeamento com menor distorção de área é escolhido.

Na etapa de suavização local da malha, a posição de cada vértice é ajustada considerando que cada triângulo de uma submalha local, consistindo dos elementos incidentes ao vértice, deve ter área proporcional à sua área na superfície original, isto é, a superfície de entrada na etapa de mapeamento inicial. Um sistema de equações lineares sobredeterminado é construído, solucionado pelo método de quadrados mínimos, e o vértice em questão é movido para a posição que apresenta menor distorção de comprimento. Já a suavização global, terceira etapa de CALD, tenta equalizar as distorções de área para todos os vértices na superfície esférica, ajustando a colatitude θ_v de cada vértice de modo a diminuir a relação entre a área de cada triângulo da malha no mapeamento e a área do triângulo correspondente na superfície original.

Além da implementação de referência de CALD chamada SPHARM-MAT,⁷⁸ escrita em MATLAB,⁷⁹ também foi experimentada implementação feita pelo estudante em linguagem C.

2.5.3 ConformalizedMCF

O terceiro tipo de mapeamento esférico foi calculado pelo programa ConformalizedMCF.⁸⁰ Utilizando uma formalização de movimento de pontos ao longo do tempo, este programa realiza uma suavização da geometria da superfície através do deslocamento de seus pontos em direção à posição média dos pontos vizinhos. A transformação converge para uma parametrização da superfície na esfera.

2.5.4 ARAP

A abordagem de mapeamento ARAP (do inglês *As Rigid As Possible*, ou tão rígido quanto possível),⁸¹ tenta preservar a rigidez da malha triangular que define a superfície e, com isso, manter a relação entre as áreas dos triângulos e minimizar as distorções. A partir de um mapeamento esférico inicial, um algoritmo iterativo de dois estágios ajusta as posições dos vértices da malha. No primeiro estágio, novas coordenadas esféricas são calculadas considerando-se efeitos locais e globais da movimentação dos vértices; no segundo estágio, o raio da esfera é ajustado para o mapeamento ser mais próximo de isométrico, ou seja, que preserve os comprimentos das arestas dos triângulos. Segundo seus autores, a escolha do mapeamento inicial não interfere nos resultados, apenas no tempo

necessário para o término do algoritmo. Assim como no método CALD, experimentações com a implementação de referência de ARAP, escrita em MATLAB, foram seguidas de implementação do algoritmo em linguagem C pelo estudante com o uso da biblioteca SuiteSparseQR⁸² para solução de sistemas de equações definidos por matrizes esparsas.

2.5.5 BLD

O mapeamento BLD (do inglês *Bijective parametrization with Low Distortion*, ou parametrização bijetiva com pequena distorção)⁸³ refina a malha na forma de solução para um problema de otimização, com penalização para degeneração de triângulos e para aumento de distorções. No entanto, BLD é tão complexo quanto ARAP e os autores não compartilharam seu código-fonte ou programa executável. O estudante, então, enviou superfícies moleculares para os autores para obter os mapeamentos BLD e, a partir destes, reconstruir as superfícies em harmônicos esféricos.

2.5.6 AHSP

Um algoritmo mais recente de mapeamento, chamado AHSP (do inglês *Advanced Hierarchical Spherical Parameterizations*, ou parametrizações esféricas hierárquicas avançadas),⁸⁴ também foi analisado, através da implementação de referência feita por seus autores. AHSP decima a malha de entrada, isto é, reduz o número de seus vértices, e simplifica regiões de elevada curvatura até a malha se tornar um tetraedro. Após trivialmente mapear o tetraedro na esfera unitária, vértices são adicionados na ordem inversa de remoção sob controle de distorções. Finalmente, uma etapa de otimização global redistribui os vértices de um modo mais uniforme pela esfera.

2.5.7 Métrica de distorção de área

A distorção de área do mapeamento é compreendida como a alteração da área de cada um dos triângulos da malha triangular original após a etapa de mapeamento esférico. Nos casos em que o triângulo t é ampliado, a distorção de área $D(t)$ é medida pela divisão de sua área $S(t)$ na esfera pela sua área original $M(t)$ na superfície molecular; a razão inversa, isto é $M(t)/S(t)$, é tomada quando o triângulo t é contraído. A equação 2.3 indica ambas as situações. A esfera é ampliada ou reduzida de modo a apresentar área total igual à área da superfície molecular. Distorção média de área $\bar{D}$ e desvio padrão σ_D , distorção máxima de área D_{max} , área esférica média $\bar{S}$, desvio padrão σ_S e área esférica máxima S_{max} foram registrados para fins de comparação entre os algoritmos de mapeamento.

$$D(t) = \begin{cases} \frac{S(t)}{M(t)}, & S(t) \geq M(t) \\ \frac{M(t)}{S(t)}, & S(t) < M(t) \end{cases} \quad (2.3)$$

2.6 Cálculo de coeficientes de harmônicos esféricos

O cálculo dos coeficientes de harmônicos esféricos pela equação 2.4 exige a determinação do valor das funções coordenadas $X(\theta, \phi)$, $Y(\theta, \phi)$ e $Z(\theta, \phi)$ para todo o domínio de valores de θ e ϕ . Não se conhece, contudo, a expressão analítica destas funções, uma vez que foram construídas a partir de uma descrição descontínua da superfície, originária das coordenadas atômicas da proteína. É necessário, portanto, o uso de algum mecanismo para estabelecer o valor da função para as posições (θ, ϕ) em que não estiver definida.

O interpolador da biblioteca Wild Magic Engine^{85,86} foi utilizado para estimar o valor das funções coordenadas. Grosso modo, o interpolador atua inicialmente subdividindo cada triângulo da malha que representa a superfície esférica em triângulos menores. Uma função interpoladora quadrática da classe C^1 , isto é, continuamente diferenciável, é então criada para receber como entrada pontos espaciais (θ, ϕ) e produzir como saída valores interpolados da função $f(\theta, \phi)$ e de suas duas derivadas parciais de primeira ordem $f_\theta(\theta, \phi)$ e $f_\phi(\theta, \phi)$. O procedimento é repetido para gerar uma função interpoladora para cada uma das três funções coordenadas.

$$\lambda_l^m = \int_0^{2\pi} \int_0^\pi F(\theta, \phi) y_l^m(\theta, \phi) \sin \theta d\theta d\phi \quad (2.4)$$

A determinação dos coeficientes $\lambda_{x,l}^m$, $\lambda_{y,l}^m$ e $\lambda_{z,l}^m$ de forma analítica não é prática porque, embora os polinômios interpoladores que aproximam as funções coordenadas tenham uma expressão simples, a avaliação dos polinômios de Legendre diretamente a partir da equação 1.3 causa instabilidade numérica para graus l elevados devido ao surgimento de números excessivamente grandes, como, por exemplo, os do termo $n!$, dificultando a obtenção da antiderivada da equação 2.4.⁸⁷ Algoritmos para solução numérica de harmônicos esféricos contornam esta dificuldade com a adoção de uma série de recorrência para cálculo dos polinômios de Legendre com aplicação de fator de normalização a cada passo.^{51,87} Uma aproximação numérica da integral da equação 2.4 com tais algoritmos foi então realizada.

Existem várias técnicas para integração numérica, com complexidade igualmente diversificada. Algumas delas são limitadas para integrais unidimensionais, impossibilitando seu uso neste trabalho. Uma técnica bastante satisfatória e que não apresenta tal limitação é a integração por Monte Carlo, um método simples e robusto baseado em números aleatórios.⁸⁸ Em probabilidade, a Lei dos Grandes Números declara que a média aritmética dos valores observados de um evento tendem, conforme se aumenta o número de observações, para o valor esperado da variável aleatória. O método de Monte Carlo utiliza esta ideia bastante intuitiva para estimar a integral de uma função $f(x)$, definida numa região S , como a média aritmética da avaliação de f em uma grande quantidade de pontos x_i , amostradas seguindo uma distribuição uniforme da variável aleatória x , conforme

definido na fórmula 2.5.⁴⁵ A função densidade de probabilidade $p(x)$ indica a probabilidade relativa de a variável aleatória x assumir um valor específico. Sua integral deve ser igual a 1, conforme definido na equação 2.6. A função densidade de probabilidade de uma distribuição uniforme na esfera unitária é igual a $\frac{1}{4\pi}$.

$$\int_S f(x)dx \approx \frac{1}{N} \sum_{i=1}^N \frac{f(x_i)}{p(x)} \quad (2.5)$$

$$\int_{-\infty}^{+\infty} p(x)dx = 1, \quad p(x) \geq 0 \quad (2.6)$$

Por trás da simplicidade do método está a necessidade de uma grande quantidade de amostras da função a ser integrada. Ainda assim, o método de Monte Carlo foi utilizado para calcular os coeficientes de harmônicos esféricos através da aproximação da equação 2.4, já que este procedimento é realizado apenas uma vez para cada função coordenada.

Os pontos de amostragem foram distribuídos uniformemente pelos triângulos da superfície esférica obtida anteriormente.⁸⁹ De modo geral, dado um triângulo esférico T com área $\mathcal{A}$, o algoritmo seleciona aleatoriamente um subtriângulo $T' \subset T$ com área $\mathcal{A}'$ uniformemente distribuída entre 0 e $\mathcal{A}$. Em seguida, um ponto é aleatoriamente escolhido sobre uma aresta do novo triângulo. A área de um triângulo esférico é dada por $\mathcal{A} = \alpha + \beta + \gamma + \pi$, em que α , β e γ são seus ângulos internos.⁹⁰ A quantidade de pontos em cada triângulo foi ajustada para ser proporcional à área $\mathcal{A}$ de cada um, resultando em maior uniformidade da distribuição. Foi utilizado o gerador de números pseudoaleatórios `rand(3)`, disponível na biblioteca padrão de funções da linguagem C.

Brechbühler, Gerig e Kübler⁴¹ sugeriram a avaliação dos coeficientes de harmônicos esféricos diretamente a partir da reconstrução da função $f(\theta, \phi)$ pela equação 2.7. Na prática, o parâmetro l é limitado por um valor arbitrário L_{max} , relacionado com a precisão desejada para a reconstrução, resultando na aproximação $f_{SH}(\theta, \phi)$ determinada pela equação 2.8. Shen e Chung⁹¹ construíram um sistema de equações lineares formado por repetições da equação 2.8, com os valores de $f(\theta, \phi)$ e de y_l^m calculados nas coordenadas (θ_v, ϕ_v) de cada vértice v do mapeamento e considerando λ_l^m como as incógnitas. O sistema de equações conseguinte é sobredeterminado e resolvido pelo método dos quadrados mínimos; neste projeto, através da rotina `dgelsy` de LAPACK.⁹² Os polinômios de Legendre da equação 1.3 necessários para obtenção dos harmônicos esféricos reais y_l^m foram calculados pela implementação de Press et al.⁵¹ e posteriormente pela biblioteca GSL.⁹³

$$f(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^l \lambda_l^m y_l^m(\theta, \phi) \quad (2.7)$$

$$f(\theta, \phi) \approx f_{SH}(\theta, \phi) = \sum_{l=0}^{L_{max}} \sum_{m=-l}^l \lambda_l^m y_l^m(\theta, \phi) \quad (2.8)$$

Reconstruções com elevado nível de detalhamento e superfícies moleculares com grande quantidade de vértices podem esgotar a memória do computador se um único sistema de equações for construído. Para permitir um uso de memória controlável, Shen e Chung⁹¹ conceberam o método IRF (*Iterative Residual Fitting*, ajuste de residual iterativo) para resolver o sistema de equações lineares para alguns valores de l de cada vez, ou seja, agrupando as bandas. A cada etapa do algoritmo, o sistema é montado e resolvido apenas para algumas bandas. A diferença entre a função original f e a reconstrução $\lambda_l^m \cdot y_l^m$, é calculada e na etapa seguinte passa a ocupar o lado esquerdo da equação 2.8. IRF explora a independência entre os harmônicos esféricos e permite combinar, em etapas individuais, harmônicos esféricos de baixo grau, que representam o aspecto geral da superfície, com harmônicos esféricos de grau elevado, que capturam seus detalhes mais finos.

2.7 Reconstrução e visualização de superfícies

Dados três conjuntos de coeficientes de harmônicos esféricos $(\Lambda_x, \Lambda_y, \Lambda_z)$, não se conhece *a priori* quais pontos do espaço compõem a superfície correspondente. Na etapa final da representação de uma superfície molecular através de harmônicos esféricos, as coordenadas tridimensionais finais da superfície são determinadas pelo cálculo da equação 2.9 para cada reconstrução de função coordenada $F_{SH} \in \{X_{SH}, Y_{SH}, Z_{SH}\}$, de modo que cada ponto i seja determinado por $(x_i, y_i, z_i) = (X_{SH}(\theta_i, \phi_i), Y_{SH}(\theta_i, \phi_i), Z_{SH}(\theta_i, \phi_i))$. O parâmetro L_{max} é ajustável de acordo com o grau de fidelidade desejado.

Os pontos de reconstrução (θ_i, ϕ_i) foram distribuídos pela esfera de três modos: a) aleatoriamente; b) periodicamente, dividindo-se o intervalo de valores válidos de cada ângulo em intervalos menores de tamanho igual; e c) seguindo uma espiral de Fibonacci na superfície da esfera⁹⁴ como uma aproximação de uma distribuição equidistante. O número de pontos a reconstruir foi ajustado para 50000, com a mesma ordem de magnitude de pontos da maioria das superfícies de teste.

Para se quantificar as diferenças entre as duas superfícies e permitir avaliação objetiva da qualidade de reconstrução, uma métrica formal se faz necessária. A raiz do desvio médio quadrático (RMSD, do inglês *root mean square deviation*) foi usada para calcular a diferença média entre as posições de pontos em superfícies diferentes. Seu valor é obtido pela equação 2.10, em que p_i é um ponto na superfície molecular original p e q_i é o ponto correspondente, isto é, mapeado para o mesmo valor de (θ, ϕ) , na superfície restaurada por harmônicos esféricos q . Porém, como as superfícies reconstruídas neste trabalho são aproximações das superfícies originais, o número de vértices pode diferir entre superfície original e reconstrução; mesmo se a quantidade de vértices for a mesma, as superfícies representadas são descritas apenas pelos coeficientes Λ_x , Λ_y e Λ_z , de modo que qualquer informação sobre o mapeamento das funções coordenadas na esfera unitária é perdida e impondo um problema para a equação 2.10. Para permitir cálculo de RMSD, e

apenas para se verificar os resultados dos métodos deste trabalho, as reconstruções foram feitas utilizando-se as mesmas coordenadas esféricas do mapeamento.

$$F_{SH}(\theta, \phi) = \sum_{l=0}^{L_{max}} \sum_{m=-l}^{m=l} \lambda_l^m y_l^m(\theta, \phi) \quad (2.9)$$

$$RMSD(p, q) = \sqrt{\frac{1}{N} \sum_{i=1}^N \|p_i - q_i\|^2} \quad (2.10)$$

Uma mera coleção de pontos no espaço tridimensional pode não ser ideal para fins de visualização, contudo. Após experimentar com o algoritmo *marching cubes*,⁷⁴ o programa Powercrust⁹⁵ foi adotado para converter a nuvem de pontos obtida da reconstrução de superfície em uma malha triangular, com conexões adequadas entre os pontos. Por fim, o programa de visualização Meshlab⁹⁶ foi usado para renderizar a malha como uma imagem.

2.8 Cálculo de curvaturas da superfície

O atracamento molecular está intimamente relacionado à superfície das moléculas participantes, e por isso é necessário identificar seu relevo. Para tanto, algumas de suas características geométricas precisam ser calculadas; futuramente, ajudarão a compor a função de pontuação que avalia quão bom é um atracamento. As curvaturas principais k_1 e k_2 determinam, respectivamente, a taxa de maior e de menor torção da superfície em um dado ponto.⁹⁷ A partir das curvaturas principais se estabelecem a curvatura gaussiana K e a curvatura média H , calculadas pelas equações 2.11 e 2.12, que caracterizam o relevo da superfície conforme indicado pela tabela 1.

$$K = k_1 k_2 \quad (2.11)$$

$$H = \frac{k_1 + k_2}{2} \quad (2.12)$$

Duncan e Olson⁴⁶ utilizaram as derivadas das funções coordenadas de superfície com relação a suas variáveis θ e ϕ para calcular as curvaturas normal e gaussiana diretamente a partir dos coeficientes de harmônicos esféricos, sem exigir portanto reconstrução da superfície. Os valores de derivada das funções coordenadas foram estabelecidos a partir da derivada dos harmônicos esféricos, como indica a equação 2.13 para X em relação a θ .

$$X_\theta = \frac{\partial X(\theta, \phi)}{\partial \theta} = \sum_{l=0}^N \sum_{m=-l}^l \lambda_{x,l}^m \frac{\partial y_l^m(\theta, \phi)}{\partial \theta} \quad (2.13)$$

O vetor normal à superfície $\hat{\mathbf{N}}$ foi definido a partir de produtos vetoriais das derivadas das funções de superfície, conforme equação 2.14, para $\mathbf{X} = (X(\theta, \phi), Y(\theta, \phi), Z(\theta, \phi))$.

Tabela 1 – Interpretação geométrica de curvaturas Gaussiana e Média em superfícies.

Curvatura	Relevo
$K > 0$	Domo
$K = 0$	Achatado em ao menos uma direção
$K < 0$	Região de sela
$H > 0$	Protuberância
$H = 0$	Vários
$H < 0$	Cavidade

Fonte: Elaborada pelo autor.

A partir de $\hat{\mathbf{N}}$, são estabelecidas as derivadas parciais $\hat{\mathbf{N}}_\theta$ e $\hat{\mathbf{N}}_\phi$. Os valores E, F, G, L, M e N , das equações 2.15 a 2.20, permitem calcular a curvatura média H pela equação 2.21 e curvatura gaussiana K pela equação 2.22.

$$\hat{\mathbf{N}} = \frac{\mathbf{X}_\theta \times \mathbf{X}_\phi}{|\mathbf{X}_\theta \times \mathbf{X}_\phi|} \quad (2.14)$$

$$E = \mathbf{X}_\theta \cdot \mathbf{X}_\theta \quad (2.15)$$

$$F = \mathbf{X}_\theta \cdot \mathbf{X}_\phi \quad (2.16)$$

$$G = \mathbf{X}_\phi \cdot \mathbf{X}_\phi \quad (2.17)$$

$$L = -\mathbf{X}_\theta \cdot \hat{\mathbf{N}}_\phi \quad (2.18)$$

$$M = \frac{1}{2}(\mathbf{X}_\theta \cdot \hat{\mathbf{N}}_\phi + \mathbf{X}_\phi \cdot \hat{\mathbf{N}}_\theta) \quad (2.19)$$

$$N = -\mathbf{X}_\phi \cdot \hat{\mathbf{N}}_\phi \quad (2.20)$$

$$H = \frac{EN + GL - 2FM}{2(EG - F^2)} \quad (2.21)$$

$$K = \frac{LN - M^2}{EG - F^2} \quad (2.22)$$

A avaliação numérica das derivadas de y_l^m , para permitir obtenção dos valores de curvatura da superfície, esbarra no surgimento de números excessivamente grandes para os polinômios de Legendre, como na seção 2.6. Para permitir o uso de algoritmos numericamente mais estáveis,^{51,87} as derivadas de y_l^m foram aproximadas numericamente pelo método das diferenças,⁵¹ como determinado pela fórmula 2.23 para derivada parcial em relação a θ .

$$\frac{\partial y_l^m(\theta, \phi)}{\partial \theta} \approx \frac{y_l^m(\theta + h, \phi) - y_l^m(\theta - h, \phi)}{2h} \quad (2.23)$$

A possibilidade de a derivação numérica pelo método das diferenças apresentar erros elevados caso o denominador $2h$ não seja escolhido adequadamente estimulou cálculo de curvaturas pela biblioteca CGAL a partir dos pontos da malha, e não dos coeficientes de harmônicos esféricos. As curvaturas principais k_1 e k_2 foram calculadas por CGAL que, após receber uma nuvem de pontos descrevendo a superfície calculada, constrói, via interpolação, uma descrição analítica própria da superfície e obtém as curvaturas principais.⁹⁸

2.9 Cálculo de energia potencial eletrostática

Outra propriedade essencial de superfícies moleculares, a energia potencial eletrostática é relacionada com a distribuição espacial de cargas elétricas e é utilizada para estimar atrações e repulsões entre as moléculas envolvidas. A equação 2.24, conhecida como equação de Poisson-Boltzmann,⁹⁹ permite calcular a distribuição de potencial elétrico $\psi(x, y, z)$ em uma superfície eletricamente carregada, dados a constante de Boltzmann k_B e os valores para a densidade local de carga c_0 , a constante dielétrica do solvente ϵ , a permissividade do espaço livre ϵ_0 e a temperatura T .

$$\nabla^2 \psi = \frac{c_0 e}{\epsilon \epsilon_0} \cdot (e^{\frac{e\psi(x,y,z)}{k_B T}} - e^{-\frac{e\psi(x,y,z)}{k_B T}}) \quad (2.24)$$

A biblioteca BALL (do inglês *Biochemical Algorithms Library*, ou biblioteca de algoritmos para bioquímica)¹⁰⁰ foi utilizada para solucionar a equação de Poisson-Boltzmann e determinar a distribuição de energia potencial eletrostática pela superfície de Connolly de uma proteína. Um arquivo PDB foi encaminhado para o módulo de BALL responsável por esta atividade, que calcula uma representação interna da superfície molecular, soluciona a equação 2.24 e, com o auxílio de uma grade cúbica, permite determinar a energia potencial eletrostática tanto na superfície quanto na vizinhança de uma proteína. Como a distribuição de energia é calculada a partir de um arquivo PDB, foi necessário representar a própria distribuição de energia potencial eletrostática em harmônicos esféricos. Assim, obtém-se um conjunto adicional de coeficientes λ_l^m que é empregado para aproximar a energia potencial eletrostática calculada por BALL.

2.10 Decomposição espectral de superfícies de um *ensemble* de conformações

Como o *ensemble* de proteínas pode ser composto de centenas de conformações, o método deste trabalho gera um conjunto contendo milhões de coeficientes. Afortunadamente, a maioria destes coeficientes será muito similar, já que, por representar conformações de uma mesma proteína, cada componente da função expandida não deve variar drasticamente. Isto é devido ao subconjunto de coeficientes respectivo à parte grosseira da representação ser o grupo dominante na forma geral da superfície. É possível, então, criar uma representação compacta onde o conjunto dominante comum entre todas as superfícies é armazenado em combinação com os conjuntos de detalhes finos para cada superfície no *ensemble*. Os conjuntos combinados devem ser capazes de criar uma hierarquia de detalhes grosseiros para finos para aproximação controlada de superfície.

O primeiro passo para se criar a representação compacta para Decomposição Espectral de Superfícies (DES) é a seleção dos coeficientes mais importantes, que devem ser capazes de reconstruir a superfície sem criar distorções significativas. Para tal fim, é necessário comparar as diferenças entre mesmos coeficientes no *ensemble*, isto é, coeficientes com mesmo índice l e mesmo índice m , e automaticamente determinar o nível de significância para as mudanças obtíveis. Os coeficientes que contribuem com alterações importantes no formato geral das estruturas do *ensemble* devem exibir mudanças significativas no seu desvio absoluto mediano (MAD, do inglês *Median Absolute Deviation*),¹⁰¹ dado pela equação 2.25. Para cada par (l, m) que descreve um coeficiente λ_l^m , há um valor $MAD(\lambda_l^m)$ associado. Na equação 2.25, $md(\chi)$ é a mediana do conjunto de dados χ , Λ é o conjunto de coeficientes λ_l^m para todas as estruturas no *ensemble*, ao passo que $\Lambda^i \in \Lambda$ é o valor de λ_l^m para a estrutura da conformação i . Em comparação com MAD, foi calculada a distância euclidiana máxima $Dist_{max}$ entre coeficientes $\Lambda^i, \Lambda^j \in \Lambda$ pela equação 2.26.

$$MAD(\lambda_l^m) = md(|\Lambda^i - md(\Lambda)|) \quad (2.25)$$

$$Dist_{max}(\lambda_l^m) = \max(|\Lambda^i - \Lambda^j|) \quad (2.26)$$

Picos nesta função indicam mudanças mais notáveis na magnitude dos coeficientes. A função que mapeia o conjunto de pares (l, m) em seus valores de $MAD(\lambda_l^m)$ é ruidosa e detecção automática de picos pode ser problemática. Se faz necessário um método capaz de reduzir os ruídos e só então detectar os picos. O método usado para esta tarefa é baseado na transformada unidimensional *wavelet* de Daubechies (1d-DWT, de *Daubechies Wavelet Transform*),¹⁰² que é um método robusto de detecção de pico quando combinado com um filtro passa-banda. Neste trabalho, a função $MAD(\lambda_l^m)$ é transformada usando a DWT e um filtro passa-banda é usado para remover valores excessivamente baixos e altos no espaço transformado. Este procedimento reduz o nível de ruído nos dados mapeados pela

função $MAD(\lambda_l^m)$. As duas etapas finais são a aplicação de DWT inversa, para trazer os dados sem ruído de volta para o espaço original, e a detecção de picos usando análise tradicional de limiar. A mais tradicional transformada unidimensional *wavelet* de Haar foi também calculada e comparada com Daubechies. O pico obtido por este procedimento deve representar o conjunto de coeficientes responsável por alterar os detalhes finos no *ensemble* conformacional.

Dados fora dos picos podem com isso ser considerados comuns entre as superfícies e usado de uma superfície de referência, enquanto os detalhes finos continuam provenientes dos demais coeficientes de cada estrutura no *ensemble*.

2.11 Conjunto de dados

Para avaliação dos diversos métodos, foram utilizados uma série de estruturas de proteínas obtidas do *Protein Data Bank*. A estrutura de 1R2M é de um homodímero de 1009 átomos, portanto constitui-se de duas moléculas idênticas formando um complexo proteico. Por ter um formato alongado, foi para fins de teste considerado como uma única superfície e empregado nos estudos de distorção de mapeamento. 2K5X_A (cadeia A da estrutura de identificação 2K5X) é a proteína colicina E9,³² uma proteína bacteriana de 816 átomos. 2LJ5 é um *ensemble* de 300 conformações de ubiquitina,²⁴ com 1231 átomos cada. Já o *Protein Docking Benchmark* reúne uma coleção de diversas estruturas determinadas experimentalmente de proteínas ligantes e receptoras e do complexo correspondente para atuar como base de dados de treinamento para os esforços de simulação de atracamento.¹⁰³ Em sua versão 5, a coleção conta com 230 complexos nas conformações *unbound* (não ligada) e *bound* (ligada), totalizando 920 estruturas. A quantidade média de átomos em suas estruturas é de 2074 ± 1788 , sendo a menor estrutura com 222 e a maior estrutura com 17454 átomos.

Além de estruturas de proteínas, as seguintes superfícies não moleculares também foram utilizadas: *armadillo*,¹⁰⁴ *cortex*³⁹ e *fandisk*.⁸¹ A tabela 2 as caracteriza em termos de quantidade de vértices, quantidade de triângulos da malha e de espaço de armazenamento ocupado.

Tabela 2 – Detalhes de estruturas selecionadas

Nome	Vértices	Triângulos	Espaço (KB)
<i>armadillo</i>	172974	345944	8109
<i>cortex</i>	40962	81920	1921
<i>fandisk</i>	608	1208	29

Fonte: Elaborada pelo autor.

3 RECONSTRUÇÃO DE SUPERFÍCIES MOLECULARES EM HARMÔNICOS ESFÉRICOS

Neste capítulo, é apresentada a avaliação da representação paramétrica de superfícies moleculares em harmônicos esféricos, referente à metodologia do capítulo 2. A seção 3.1 inicia a exposição de resultados com a verificação de ortonormalidade das funções, para confirmar que, de fato, formam a base de um espaço de funções. As várias combinações de algoritmos para se obter a representação final são avaliadas na seção 3.2, seguidas por discussões acerca de distorções do mapeamento esférico na seção 3.3 e sua influência na qualidade da restauração das funções projetadas. Na seção 3.4 são mostrados os resultados de fechamento de túneis na superfície, ao passo que na seção 3.5 são indicadas as reconstruções de um conjunto numeroso de estruturas. As curvaturas da superfície e a energia potencial eletrostática obtidas para uma estrutura são apresentadas nas seções 3.6 e 3.7, respectivamente. Por fim, a seção 3.8 discute os resultados de decomposição espectral.

Para auxiliar a compreensão, os resultados da seção 3.2 são organizados em ciclos completos de cálculo de superfícies, mapeamento esférico, cálculo de coeficientes e reconstrução.

3.1 Ortonormalidade de harmônicos esféricos

Para estabelecer confiança na implementação de harmônicos esféricos, comparou-se a implementação destas funções deste projeto com a dos programas estabelecidos MATLAB⁷⁹ e Mathematica.¹⁰⁵ O erro absoluto obtido foi inferior a 10^{-6} . Seguiu-se então a avaliação da equação 2.1 para ortonormalidade de harmônicos esféricos. A figura 2 mostra os valores calculados para $l = l', m = m'$ variando-se o parâmetro l de 0 a 10. Como esperado, o valor médio calculado para $\delta_{ll'}\delta_{mm'}$ foi igual a 1,000. Já a figura 3 mostra os valores para $l \neq l', m \neq m'$ para $l \leq 5$, que resultaram no valor médio de 0,000. O índice de harmônicos esféricos é medido consecutivamente, começando em $(l, l', m, m') = (0, 0, 0, 0)$ e terminando em $(l, l', m, m') = (10, 10, 10, 10)$ no primeiro caso e em $(l, l', m, m') = (5, 5, 5, 5)$ no segundo caso. A pequena variação em torno dos valores 1 e 0 confirmou, como esperado devido à ortogonalidade e normalização dos harmônicos esféricos, o valor 1 para harmônicos esféricos iguais e o valor zero para harmônicos esféricos diferentes. A integração também foi testada substituindo-se os harmônicos esféricos da equação 2.1 pelas as expressões analíticas de Gross e Staniforth.¹⁰⁶ Os coeficientes calculados pela integral para $\delta_{ll'}\delta_{mm'}$ foram nulos para $l \neq l'$ e $m \neq m'$ e iguais a aproximadamente 0,7 para os demais casos, devido a outro fator de normalização utilizado, portanto aprovando a integração numérica.

Figura 2 – Valores calculados de $\delta_{ll'}\delta_{mm'}$ para harmônicos esféricos com índices iguais $l = l'$, $m = m'$ até $l = 10$.

Fonte: Elaborada pelo autor.

3.2 Reconstrução de superfícies em harmônicos esféricos

3.2.1 Surfmap e integração numérica

O estudo se iniciou com o cálculo de superfícies moleculares (MS ou Connolly) pelo programa Surfmap para o complexo de código PDB 1R2M. A malha triangular gerada por Surfmap nas configurações padrão, contendo 2026 vértices, é mostrada na figura 4 (a). A quantidade máxima de vértices obtida para esta molécula nesta ferramenta foi de 7204, ajustando-se o parâmetro não documentado *fill* de 50% (padrão) para 74%. Acima deste valor, mais de uma superfície é gerada e o programa não consegue gerar um mapeamento na esfera. A malha mais detalhada resultante é mostrada na figura 4 (c).

Para calcular os coeficientes de harmônicos esféricos por integração numérica, foram gerados cem mil pontos aleatórios na superfície esférica para amostragem, distribuídos pelos triângulos de sua malha.⁸⁹ Como há milhares de triângulos, o tamanho de muitos deles é suficientemente pequeno para torná-los compatíveis com triângulos planares, cuja soma dos ângulos internos é igual a π rad. Devido à precisão finita dos cálculos, a área triangular $\mathcal{A}$ resultava em zero. A substituição da área de triângulo esférico pela área de triângulo planar resolveu a questão. A figura 5 mostra a distribuição de 10 pontos pela área de um único triângulo da malha. Por inspeção visual, a distribuição de pontos apresentou cobertura adequada da superfície esférica.

Após o cálculo dos coeficientes de harmônicos esféricos, a superfície original foi

Figura 3 – Valores calculados de $\delta_{ll'}\delta_{mm'}$ para harmônicos esféricos com índices diferentes $l \neq l'$, $m \neq m'$ até $l = 5$.

Fonte: Elaborada pelo autor.

Figura 4 – Malhas calculadas pelo programa Surfmap para o complexo 1R2M: (a) superfície molecular com 2026 vértices; (b) mapeamento esférico com 2026 vértices; (c) superfície molecular com 7204 vértices; (d) mapeamento esférico com 7204 vértices.

Fonte: Elaborada pelo autor.

Figura 5 – Distribuição aleatória de 10 pontos (esferas escuras) na superfície de um triângulo esférico (com esferas claras como vértices).

Fonte: Elaborada pelo autor.

Figura 6 – Reconstrução da superfície molecular de 1R2M utilizando o mapeamento de Surfmap e integração numérica com (a) 2026 vértices e $L_{max} = 20$; (b) 2026 vértices e $L_{max} = 40$; (c) 7204 vértices e $L_{max} = 20$, (d) 7204 vértices e $L_{max} = 30$ e (e) 7204 vértices e $L_{max} = 40$.

Fonte: Elaborada pelo autor.

reconstruída aproximadamente pelo cálculo das expressões 2.8 para as funções X, Y e Z. As reconstruções obtidas para a superfície de 2026 vértices são mostradas na figura 6 (a) com $L_{max} = 20$ e na figura 6 (b) com $L_{max} = 40$; já as reconstruções para a superfície de 7204 vértices são mostradas nas figuras 6 (c-e) com $L_{max} = 20, 30$ e 40 , respectivamente. Enquanto a restauração com menor valor de L_{max} apresenta insuficiência de detalhes, as restaurações com L_{max} maior, que se esperava aumentar o nível de detalhamento e tornar-se mais similares à superfície original, manifestaram distorções preocupantes, refletivas no aumento de RMSD com o aumento de bandas, de 1,3046 Å para 1,8432 Å (2026 vértices), e de 3,5456 Å para 4,8467 Å (7204 vértices).

Embora o mapeamento esférico esteja correto, Surfmap não conseguiu gerar o mapeamento para outras moléculas além de 1R2M, incluída como exemplo em seu pacote, e foi incapaz até mesmo de mapear a mesma superfície molecular de 1R2M após rotação de 90 graus em torno de qualquer um dos eixos coordenados, impossibilitando a continuidade de seu uso neste estudo. Diante disso, a ferramenta foi abandonada em favor da combinação das bibliotecas ESBTL e CGAL.

Suspeitou-se de que as distorções na superfície reconstruída fossem causadas pelo interpolador. Testes considerando o interpolador isoladamente identificaram instabilidade numérica na interpolação nos polos e nas regiões circumpolares, que, combinada com os resultados satisfatórios da integração numérica de expressões analíticas da seção 3.1, condenaram o interpolador utilizado. O interpolador de Gross e Staniforth¹⁰⁶ foi consi-

Figura 7 – Malhas referentes à superfície de 1R2M gerada por CGAL com (a) 8167 vértices, (b) 21445 vértices e (c) 50333 vértices.

Fonte: Elaborada pelo autor.

derado como alternativa, mas se mostrou pouco flexível. A dificuldade de incorporá-lo à metodologia deste trabalho causou sua rejeição.

3.2.2 ESBTL, CGAL, CALD e IRF

As triangulações de 1R2M resultantes da combinação de ESBTL⁷⁰ e CGAL⁷⁶ são mostradas na figura 7 para malhas com 8167, 21445 e 50333 vértices. É notável a irregularidade na distribuição de vértices pela malha triangular que forma a superfície e menor detalhamento em comparação às malhas de Surfmap com quantidade similar de vértices.

As três etapas do algoritmo de mapeamento esférico CALD⁴⁰ foram aplicadas na malha da superfície de 1R2M calculada por CGAL da figura 7 (b), resultando na malha da figura 8 (a). Ao contrário das baixas distorções prometidas pelo algoritmo, o aspecto visual do mapeamento piorou em comparação ao de Surfmap, se tornando menos organizado, e o detalhe da figura 8 (b) mostra que alguns triângulos da malha se sobrepuseram aos vizinhos. Sobreposições de triângulos são indesejadas por alterarem a proporção entre as áreas de triângulos vizinhos, em comparação à proporção de áreas de triângulos na superfície que originou o mapeamento, provocando aumento na distorção de área do mapeamento. Nota-se também baixa concentração de vértices nos polos em relação à região equatorial, aumentando ainda mais a distorção de área do mapeamento. Neste algoritmo foram encontradas dificuldades relacionadas à precisão finita da representação de números reais pelo computador. Como há elevada quantidade de triângulos na malha que define a superfície esférica, alguns deles são tão pequenos que se aproximam de triângulos planares, fazendo com que sua área, calculada como a soma dos ângulos internos subtraída de π , resultasse em 0 e prejudicando a comparação entre os três mapeamentos intermediários calculados pelo algoritmo. A solução encontrada foi ampliar a superfície molecular original através da multiplicação de cada um de seus pontos por um fator de escala até que todos os triângulos apresentassem área não nula. O raio da superfície esférica também foi aumentado através da equação 3.1, onde R_{novo} é o raio ajustado e A é a área

Figura 8 – (a) Mapeamento obtido com CALD para a superfície de 21445 vértices gerada por CGAL para o complexo 1R2M; (b) detalhe do mapeamento, com destaque para triângulos superpostos.

Fonte: Elaborada pelo autor.

da superfície molecular, para que as áreas das duas superfícies coincidisse. Ainda assim, alguns triângulos muito pequenos continuaram com área esférica nula, causada também por questões de precisão ao se calcular a distância esférica entre dois de seus vértices. Optou-se por ignorar tais triângulos devido à sua pequena contribuição na métrica de distorção de área.

$$R_{novo} = \sqrt{\frac{A}{4\pi}} \quad (3.1)$$

O efeito do aumento da quantidade de vértices na superfície original foi estudado com a reconstrução de duas superfícies de 1R2M, com 3557 vértices e com 7114 vértices, calculadas novamente com ESBTL e CGAL. Os respectivos mapeamentos, calculados por CALD, apresentam as mesmas características de sobreposição de triângulos e de diferenças de concentração de vértices mencionados previamente. A figura 9 ilustra a reconstrução, utilizando uma versão inicial do algoritmo IRF para cálculo dos coeficientes de harmônicos esféricos, da superfície de 3557 pontos de 1R2M com (a) $L_{max} = 20$ e (b) $L_{max} = 30$. Em (a) há pequenas distorções na parte inferior da superfície, ao passo que em (b) as distorções tornam a superfície irreconhecível. A figura 10 mostra a reconstrução da superfície contendo 7114 vértices nas situações com (a) $L_{max} = 20$, (b) $L_{max} = 30$ e (c) $L_{max} = 35$. O aumento da quantidade de vértices resultou em melhor qualidade de representação; no entanto ainda se manifestaram distorções acentuadas quando o parâmetro L_{max} foi aumentado para 35.

A superfície da figura 7 foi também submetida à implementação de referência de CALD, SPHARM-MAT, originando o mapeamento da figura 11 (a) e (b). Não foram detectadas sobreposições de triângulos; no entanto, houve grande concentração de vértices em quatro regiões paralelas ao equador.

Figura 9 – Reconstrução, utilizando as implementações de CALD e IRF deste projeto, da superfície molecular de 1R2M com 3557 vértices e (a) $L_{max} = 20$ e (b) $L_{max} = 30$.

Fonte: Elaborada pelo autor.

Figura 10 – Reconstrução, utilizando as implementações de CALD e IRF deste projeto, da superfície molecular de 1R2M com 7114 vértices e (a) $L_{max} = 20$, (b) $L_{max} = 30$ e (c) $L_{max} = 35$.

Fonte: Elaborada pelo autor.

Figura 11 – Mapeamento esférico da superfície do complexo 1R2M calculado pelo programa SPHARM-MAT (a) e mesmo mapeamento após rotação de 90 graus em torno do eixo horizontal (b).

Fonte: Elaborada pelo autor.

Figura 12 – Mapeamento da superfície de 1R2M, calculado por ConformalizedMCF, utilizando a triangulação de CGAL: (a) com 21445 vértices; (b) com 50333 vértices.

Fonte: Elaborada pelo autor.

Figura 13 – Reconstrução da superfície de 1R2M utilizando a triangulação com 21445 vértices de CGAL e mapeamento de ConformalizedMCF: (a) $L_{max} = 20$; (b) $L_{max} = 30$; (c) $L_{max} = 40$.

Fonte: Elaborada pelo autor.

3.2.3 ESBTL, CGAL, ConformalizedMCF e IRF

Dada a impossibilidade de reprodução dos resultados de CALD, os mapeamentos esféricos foram então calculados pelo ConformalizedMCF⁸⁰ a partir das superfícies de 1R2M de CGAL com 21445 vértices e 50333 vértices. Nas malhas resultantes, figura 12, nota-se grande diferença de área entre os triângulos que formam cada uma das malhas, diferença esta que também se observa na triangulação da superfície molecular. Não obstante, não foram observadas sobreposições de triângulos ou elevadas concentrações em algumas poucas regiões da esfera, como ocorreu previamente com CALD.

A reconstrução em harmônicos esféricos da superfície com 21445 vértices é mostrada na figura 13 para (a) $L_{max} = 20$, (b) $L_{max} = 30$ e (c) $L_{max} = 40$. As distorções observadas anteriormente não se manifestaram, o que pode ser atribuído ao novo algoritmo de mapeamento, mas também à maior quantidade de vértices na superfície. No entanto, a observação da figura 14, contendo a reconstrução da superfície com 50333 pontos, mais do que o dobro do que a superfície anterior, para os mesmos valores de L_{max} , mostra que o acréscimo de vértices por si só não bastou para conter as distorções.

Figura 14 – Reconstrução da superfície de 1R2M utilizando a triangulação com 50333 vértices de CGAL e mapeamento de ConformalizedMCF: (a) $L_{max} = 20$; (b) $L_{max} = 30$; (c) $L_{max} = 40$.

Fonte: Elaborada pelo autor.

Figura 15 – Superfície *cortex*: (a) original; (b) mapeamento ConformalizedMCF; (c) reconstrução com $L_{max} = 100$.

Fonte: Elaborada pelo autor.

3.2.4 Triangulação pronta e IRF

Tornou-se necessária uma abordagem diferente para a eliminação dos erros. Para tanto, optou-se por omitir o estágio de triangulação inicial e alimentar o programa com uma superfície previamente triangulada. Experimentou-se, primeiramente, com a superfície *cortex*,³⁹ figura 15 (a), que possui grande quantidade de cavidades. *Cortex* é uma superfície já triangulada e exibe alta regularidade na distribuição de triângulos, isto é, pequena diferença de área entre os triângulos da malha. No mapeamento resultante não houve sobreposição de triângulos ou diferenças gritantes de concentração de vértices, como ilustrado pela figura 15 (b). A reconstrução de *cortex* com o algoritmo IRF e $L_{max} = 100$, mostrada na figura 15, apresenta aparente baixo nível de distorção e elevada representatividade do aspecto geral da superfície original, ainda que detalhes finos tenham sido omitidos.

Os testes de mapeamento e reconstrução prosseguiram com a superfície *armadillo*,¹⁰⁴

Figura 16 – (a) superfície *armadillo*; (b) e (c) diferentes orientações do mapeamento de *armadillo* calculado por SPHARM-MAT; (d) mapeamento de *armadillo* calculado por ConformalizedMCF.

Fonte: Elaborada pelo autor.

também já triangulada e mostrada na figura 16 (a). Para fins de comparação, *armadillo* foi submetido ao programa SPHARM-MAT, que gerou o mapeamento das figuras 16 (b) e (c), sem sobreposição de triângulos, porém mantendo a tendência de concentrar vértices em anéis paralelos, o que não foi observado no mapeamento calculado por ConformalizedMCF da figura 16 (d). A reconstrução final calculada por SPHARM-MAT com $L_{max} = 50$ é mostrada pela figura 17 (a) e apresenta ausência de diversas partes componentes da superfície original. O mapeamento de SPHARM-MAT foi ainda utilizado para o cálculo dos coeficientes de harmônicos esféricos com $L_{max} = 50$ pela implementação de IRF deste trabalho, culminando com a superfície da figura 17 (b). Novamente, há ausência de partes da superfície original, substituídas por distorções grosseiras. Também se manifestou acentuada rugosidade. A reconstrução final, utilizando o mapeamento de ConformalizedMCF da figura 16 (c), com $L_{max} = 50$, é mostrada na figura 17 (c). O aspecto visual da reconstrução é facilmente correspondido com o da superfície original e a ausência de detalhes mais finos é compatível com o truncamento representado por L_{max} e o elevadíssimo nível de detalhes da superfície. Não obstante, não se manifestaram cortes na superfície.

3.2.5 EDTSurf, ConformalizedMCF e IRF

Uma vez que *cortex* e *armadillo* permitiram reconstruções sem distorções excessivas, e como suas malhas triangulares originais apresentam triangulação mais uniforme do que superfícies calculadas por CGAL, investigou-se a possibilidade de a distribuição menos uniforme de triângulos pela superfície original contribuir, de algum modo, para piores resultados na reconstrução. A figura 18 mostra superfícies de Connolly de 1R2M, 1R2M_B e 2K5X_A calculadas pelo programa EDTSurf.⁷³ O número de vértices foi de, respectivamente, 152364, 94226 e 76273. Para as 920 estruturas de *Protein Docking Benchmark*, a quantidade média de vértices da malha resultante foi de 115893 ± 53204 , com mínima de 22259 e máxima de 276514 vértices. Há notório maior detalhamento e maior regularidade das áreas dos triângulos da malha em comparação com superfícies geradas

Figura 17 – Reconstruções da superfície *armadillo* com $L_{max} = 50$: (a) calculada por SPHARM-MAT; (b) calculada pela implementação de IRF deste projeto a partir do mapeamento de SPHARM-MAT; (c) calculado também por IRF a partir do mapeamento de ConformalizedMCF.

Fonte: Elaborada pelo autor.

Figura 18 – Superfícies de Connolly calculadas por EDTSurf: (a) de 1R2M; (b) de 1R2M_B; (c) de 2K5X_A.

Fonte: Elaborada pelo autor.

por CGAL. Os mapeamentos esféricos resultantes, calculados por ConformalizedMCF, são mostrados na figura 19.

A triangulação de 1R2M calculada por EDTSurf foi reconstruída com IRF e originou a superfície de Connolly mostrada na figura 20 (a) para $L_{max} = 50$ e RMSD de cerca de 0,475 Å. Há acentuada rugosidade na região central da superfície, que engloba a interface entre os dois protômeros do complexo. A rugosidade se intensificou com o aumento de L_{max} . Uma comparação direta entre sua função X original e a reconstrução X_{SH} é mostrada na figura 21; ainda que haja elevada correspondência geral entre os dois conjuntos de valores, com correlação superior a 99%, alguns de seus valores divergem sensivelmente do esperado. Já a reconstrução da superfície de Connolly de 1R2M_B com $L_{max} = 50$, mostrada na figura 20 (b), apresentou RMSD de cerca de 0,08 Å e elevada correspondência visual entre a superfície original e a reconstrução. A reconstrução de 2K5X_A, também com $L_{max} = 50$, ilustrada pela figura 20 (c), resultou em RMSD inferior a 0,47 Å e com rugosidades similares às de 1R2M e arredondamento nas extremidades. Nos três casos, o aumento de L_{max} não eliminou o arredondamento e acentuou as rugosidades.

Figura 19 – Mapeamentos calculados por ConformalizedMCF: (a) de 1R2M; (b) de 1R2M_B; (c) de 2K5X_A.

Fonte: Elaborada pelo autor.

Figura 20 – Reconstruções das superfícies da figura 18 com $L_{max} = 50$: (a) de 1R2M; (b) de 1R2M_B; (c) de 2K5X_A.

Fonte: Elaborada pelo autor.

Ao se examinar os mapeamentos da figura 19 da superfície de Connolly das estruturas moleculares utilizadas, nota-se que em todos os casos há regiões com elevada concentração de vértices e regiões com concentração reduzida, principalmente em 1R2M, coincidentemente a estrutura cuja reconstrução manifestou rugosidades mais destacadas. O trabalho de Brechbühler, Gerig e Kübler⁴¹ já alertara para os efeitos prejudiciais de um mapeamento não uniforme, em que a densidade de vértices (e, por extensão, de triângulos) na superfície da esfera não é homogênea. Ainda no mesmo trabalho, durante o cálculo de coeficientes de harmônicos esféricos, algumas regiões da esfera ficariam representadas demais, devido à maior presença de vértices, lesando as regiões com menor concentração e resultando em coeficientes distorcidos. Redução nas quantidades de vértices da malha não produziu redução notável nestas distorções.

3.2.6 ARAP e IRF

As experimentações com mapeamento migraram para o programa ARAP (*As Rigid As Possible*, ou tão rígido quanto possível) que se dispõe a preservar a rigidez da malha triangular da superfície de entrada. A partir da superfície de exemplo *fandisk*, fornecida com

Figura 21 – Função coordenada original X de 1R2M e sua reconstrução X_{SH} usando o mapeamento de ConformalizedMCF e $L_{max} = 50$.

Fonte: Elaborada pelo autor.

Figura 22 – Superfície *fandisk*: (a) original; (b) reconstrução a partir do mapeamento esférico de ConformalizedMCF e $L_{max}=22$; (c) reconstrução a partir do mapeamento esférico ARAP e $L_{max} = 22$.

Fonte: Elaborada pelo autor.

Figura 23 – Mapeamentos da superfície *fandisk*: (a) de ConformalizedMCF; (b) de ARAP.

Fonte: Elaborada pelo autor.

o programa e mostrada na figura 22 (a), obteve-se o mapeamento de ConformalizedMCF da figura 23 (a) e o mapeamento ARAP da figura 23 (b). O mapeamento ARAP também gerou regiões com diferentes densidades de triângulos, mas a diferença de concentração não é tão acentuada quanto o de ConformalizedMCF.

A reconstrução de *fandisk* a partir de cada um dos dois mapeamentos anteriores é apresentada na figura 22 (b) e 22 (c) para $L_{max} = 22$. Ambas possuem RMSD baixo em relação à superfície original, mas há claramente uma melhor correspondência da reconstrução a partir do mapeamento ARAP com a superfície original. Valores superiores de L_{max} não puderam ser usados devido à baixa quantidade de vértices de *fandisk*, que impediu a formação de um sistema de equações sobredeterminado e portanto inibiu sua solução pelo método de quadrados mínimos de IRF.

A implementação de referência de ARAP, feita no software MATLAB que usa linguagem de programação interpretada e fracamente tipada, levou tempo excessivamente alto para tentar mapear superfícies moleculares, ainda maior do que SPHARM-MAT, devido à maior complexidade das operações. A implementação do mapeamento esférico ARAP escrita pelo aluno em linguagem de programação C foi bem-sucedida nas superfícies de exemplo disponíveis com a implementação de referência e apresentou desempenho superior. No entanto, não foi possível mapear superfícies moleculares. Ao contrário do que ocorreu com as superfícies de exemplo, números complexos surgiram em resultados intermediários ao se aplicar o algoritmo nas superfícies moleculares testadas, exigindo tratamento adequado na implementação em linguagem C e que é feito automaticamente pelo programa MATLAB. Os autores de ARAP, após serem contactados, recomendaram o uso de seu trabalho mais recente, BLD.

Figura 24 – Mapeamentos esféricos de (a) 1R2M e de (b) 2K5X obtidos com BLD.

Fonte: Elaborada pelo autor.

3.2.7 EDTSurf, BLD e IRF

As figuras 24 (a) e 24 (b) mostram o mapeamento BLD das superfícies moleculares de 1R2M e 2K5X_A feitas pelos autores de BLD a pedido do aluno. O código-fonte deste método não foi compartilhado com o aluno, com a alegação de uso de bibliotecas de software proprietárias (com, possivelmente, termos de licença de uso que dificultariam o compartilhamento do código). Em comparação com ConformalizedMCF, o mapeamento BLD é mais regular em sua distribuição de triângulos, não apresentando as grandes diferenças de concentração de triângulos observadas no mapeamento de ConformalizedMCF e associadas com distorções na reconstrução da superfície.

Os mapeamentos calculados pelo método BLD foram utilizados para reconstruir as superfícies correspondentes. A reconstrução em harmônicos esféricos com IRF da superfície de 1R2M é mostrada na figura 25 para (a) $L_{max} = 20$, (b) $L_{max} = 50$ e (c) $L_{max} = 77$. Maiores valores de L_{max} não puderam ser utilizados pois geraram resultados errôneos compatíveis com instabilidade numérica. O aumento progressivo de detalhamento foi adequado e não surgiram rugosidades como nas reconstruções com outros mapeamentos. Resultados igualmente favoráveis foram observados para a reconstrução da superfície de 2K5X_A, como pode ser visto na figura 26 para (a) $L_{max} = 20$, (b) $L_{max} = 50$ e (c) $L_{max} = 100$. Além de elevada correspondência visual, arredondamentos e rugosidades não se manifestaram. A correlação entre a função original X de 1R2M e sua reconstrução X_{SH} com $L_{max} = 50$ foi superior a 99%, mas ao contrário da reconstrução a partir do mapeamento ConformalizedMCF, nesta a correspondência foi superior, o que não significa, contudo, que X_{SH} tenha sido igual a X , apenas que as distorções observadas anteriormente não foram observadas aqui.

3.2.8 EDTSurf, AHSP e IRF

O algoritmo *Advanced Hierarchical Spherical Parameterizations* (AHSP), disponível na implementação de referência de seus autores, foi usado para mapear as superfícies

Figura 25 – Reconstrução de 1R2M usando o mapeamento de BLD e (a) $L_{max} = 20$, (b) $L_{max} = 50$ e (c) $L_{max} = 77$.

Fonte: Elaborada pelo autor.

Figura 26 – Reconstrução de 2K5X_A usando o mapeamento de BLD e (a) $L_{max} = 20$, (b) $L_{max} = 50$ e (c) $L_{max} = 100$.

Fonte: Elaborada pelo autor.

de 1R2M e de 2K5X_A, resultando nos mapas esféricos das figuras 28 (a) e 28 (b). É pertinente destacar que em cada um destes mapeamentos a distribuição de vértices pela superfície esférica não é homogênea, embora a diferença de concentração seja menor do que nos mapeamentos ConformalizedMCF. A comparação da figura 29 identifica elevada correlação, superior a 99%, entre os valores de X e os valores reconstruídos X_{SH} com $L_{max} = 50$. Neste exemplo a correspondência de valores foi ligeiramente inferior ao exemplo com mapeamento BLD, mas valores excessivamente destoantes não foram detectados na figura 29, o que se refletiu na acentuada correspondência das superfícies reconstruídas das figuras 30 (a) e 30 (b), ambas com $L_{max} = 50$, com as respectivas superfícies originais. A seção 3.3 a seguir discorre sobre consequências da distribuição de vértices no mapeamento esférico.

3.3 Distorções do mapeamento

As tabelas 3 e 4 contêm as distorções de área da equação 2.3 calculadas para os mapeamentos ConformalizedMCF, BLD e AHSP para 1R2M e para 2K5X_A, respectivamente. O mapeamento AHSP resultou em triângulos que são em média menos distorcidos (coluna $\bar{D}$) do que os de ConformalizedMCF. Tal distorção foi mais consistente, refletindo-

Figura 27 – Função coordenada original X de 1R2M e sua reconstrução X_{SH} usando o mapeamento de BLD e $L_{max} = 50$.

Fonte: Elaborada pelo autor.

Figura 28 – Mapeamentos esféricos calculados por AHSP: (a) da superfície de 1R2M; (b) da superfície de 2K5X_A.

Fonte: Elaborada pelo autor.

Figura 29 – Função coordenada original X de 1R2M e sua reconstrução X_{SH} usando o mapeamento de AHSP e $L_{max} = 50$.

Fonte: Elaborada pelo autor.

Figura 30 – Reconstruções a partir de mapeamentos AHSP: (a) superfície de 1R2M com $L_{max} = 50$; (b) superfície de 2K5X_A com $L_{max} = 50$.

Fonte: Elaborada pelo autor.

Tabela 3 – Métricas de distorção para os mapeamentos ConformalizedMCF, BLD e AHSP da superfície de 1R2M.

Algoritmo	$\bar{D}$	σ_D	D_{max}	$\bar{S}$	σ_S	S_{max}
ConformalizedMCF	8.6556	12.8469	225.2050	0.0374	0.0692	1.6175
BLD	1.7169	0.4701	3.3194	0.0374	0.0239	0.1673
AHSP	1.7051	0.4477	3.7948	0.0374	0.0258	0.1958

Fonte: Elaborada pelo autor.

Tabela 4 – Métricas de distorção para os mapeamentos ConformalizedMCF, BLD e AHSP da superfície de 2K5X_A.

Algoritmo	$\bar{D}$	σ_D	D_{max}	$\bar{S}$	σ_S	S_{max}
ConformalizedMCF	19.9622	147.9880	4974.4300	0.0378	0.0302	0.3203
BLD	1.4080	0.4013	3.6910	0.0378	0.0156	0.1443
AHSP	1.5054	0.4066	3.4773	0.0378	0.0180	0.1358

Fonte: Elaborada pelo autor.

se em menor desvio padrão de distorção σ_D . Ademais, as áreas triangulares em AHSP variam menos (menor desvio padrão de área triangular σ_S) do que em ConformalizedMCF, denotando que suas áreas são mais similares entre si e portanto a distribuição de triângulos é mais uniforme na superfície da esfera, evitando regiões com concentração excessivamente baixa de pontos. AHSP resultou, ainda, em menor distorção máxima de área D_{max} e menor área máxima de triângulo S_{max} do que ConformalizedMCF. O mapeamento BLD, todavia, apresentou resultados iguais ou superiores nas métricas e nos exemplos avaliados. As reconstruções utilizando AHSP, indicadas na figura 30, são equivalentes em qualidade visual às do mapeamento BLD, com baixo grau de distorções, ainda que em ligeira desvantagem em relação a este. Embora Duncan e Olson⁴⁶ não comentem sobre densidade de pontos na superfície esférica, este resultado confirma Brechbühler, Gerig e Kübler⁴¹ acerca da necessidade de se atingir ampla cobertura na esfera ao se projetar funções em harmônicos esféricos.

A figura 31 (a) mostra a distorção de área do mapeamento ConformalizedMCF da superfície de 1R2M, ao passo que a figura 31 (b) mostra a distância euclidiana entre cada ponto na superfície original e o ponto correspondente reconstruído, i.e. reconstruído utilizando a mesma coordenada (θ, ϕ) . Observa-se que regiões do mapeamento com maior distorção de área, em amarelo e vermelho na figura 31 (a), correspondem a regiões com maior distância euclidiana na reconstrução, também em amarelo e vermelho na figura 31 (b), implicando, assim, em uma associação entre distorções no mapeamento e distorções na reconstrução. Contudo, a mesma associação não é evidente entre o mapeamento AHSP

de 1R2M da figura 32 (a) e a respectiva reconstrução na figura 32 (b). Por outro lado, as magnitudes da distorção de área e da distorção de reconstrução com AHSP foram ao menos uma ordem de grandeza menores do que com ConformalizedMCF. Brechbühler, Gerig e Kübler⁴¹ comentam que, ao se calcular os coeficientes de harmônicos esféricos utilizando um mapeamento não homogêneo, se “espera uma sobre-representação de algumas partes em detrimento de outras, resultando em descritores distorcidos” [p. 163, tradução nossa], no entanto sem especificar em quais posições (θ, ϕ) seus descritores ficaram distorcidos. Uma análise matemática rigorosa da relação entre elevada distorção de área no mapa e de alta distorção na reconstrução é sugerida como um trabalho futuro.

Ainda que AHSP gere mapeamentos não regulares e portanto com variação da distribuição de vértices, reconstruções com este mapeamento apresentaram qualidade superior às feitas a partir dos demais mapeamentos, exceto BLD a que não se teve acesso. Isto é compatível com os menores valores de distorção de área observados em AHSP e com menor diferença de densidade de vértices de AHSP em relação a ConformalizedMCF e CALD.

Conclui-se que, para obtenção de distribuição regular de pontos na superfície esférica e consequente maior qualidade nas reconstruções, é necessário um algoritmo de mapeamento que preserve ao máximo possível a relação entre a área de cada triângulo da malha original e a área total da malha, combinado com um algoritmo de cálculo de superfície molecular e criação de malha que tenha distribuição regular de triângulos. A distribuição regular na malha triangular da superfície original garante captura mais ampla de valores da função de interesse, seja ela as funções de formato X , Y ou Z , ou outras funções definidas na superfície, como, por exemplo, energia potencial eletrostática. A preservação de área do mapeamento tem duplo propósito: (1) fixar população mais ampla do domínio esférico, estabelecendo melhor cobertura dos parâmetros (θ, ϕ) ; e (2) transferir a regularidade de distribuição de valores da função representada para o domínio esférico. Apenas a regularidade na distribuição de pontos na malha original não é suficiente para obtenção de resultados de alta qualidade, como foi o caso da combinação de EDTSurf e ConformalizedMCF. Por outro lado, nem a preservação de áreas no mapeamento nem a distribuição homogênea de pontos no domínio esférico garantem por si só a captura ampla de valores da função representada. EDTSurf e AHSP atenderam a estes requisitos para as superfícies examinadas.

3.4 Detecção e fechamento de túneis

A figura 33 mostra que, dada uma malha com túnel como entrada (a), o programa ReebHanTun corretamente identifica os vértices pertencentes ao túnel (b), que podem então ser removidos, resultando num buraco na malha (c); por fim, CGAL fecha o túnel (d). Este processo é completamente automático e é repetido até não restar túneis na malha

Figura 31 – (a) Distorção de área $D(t)$ do mapeamento ConformalizedMCF de 1R2M; (b) distância euclidiana entre pontos originais de 1R2M e respectivos pontos reconstruídos com $L_{max} = 50$.

Fonte: Elaborada pelo autor.

Figura 32 – (a) Distorção de área $D(t)$ do mapeamento AHSP de 1R2M; (b) distância euclidiana entre pontos originais de 1R2M e respectivos pontos reconstruídos com $L_{max} = 50$.

Fonte: Elaborada pelo autor.

do objeto. A detecção de túneis não é livre de falhas, no entanto. Das 920 estruturas da coleção *Protein Docking Benchmark*, por exemplo, 512 possuem *genus* maior do que zero. ReebHanTun falhou em detectar túneis em sete delas, ou 1,367% dos casos com túneis. Esta baixa taxa de falha pode ser contornada com edição manual da malha para remoção de túneis através de programas estabelecidos, como, por exemplo, Meshlab,⁹⁶ Blender¹⁰⁷ e Meshmixer.¹⁰⁸

O método proposto neste trabalho é, além de um aperfeiçoamento, uma generalização da abordagem de representação de superfícies de Duncan e Olson⁴⁶ para superfícies de qualquer *genus*. No entanto, a eliminação de túneis desse modo, com a adição de uma camada de vértices que atua como parede, introduz artificialmente uma região côncava na superfície e, sendo um acidente em seu relevo, impõe dificuldade adicional no mapeamento esférico. Como consequência, túneis na superfície original potencializam distorções na superfície representada em harmônicos esféricos. Por exemplo, as estruturas de *Protein Docking Benchmark* de *genus* zero quando reconstruídas com $L_{max} = 50$ foram, em média, menos distorcidas do que estruturas da mesma coleção com *genus* maior do que zero reconstruídas com mesmo L_{max} ; o RMSD médio de reconstruções do primeiro grupo é de $0,0961 \pm 0,0536$ Å, ao passo que o RMSD médio de reconstruções do segundo grupo, de *genus* maior do que zero, é de $0,2113 \pm 0,1290$ Å. No entanto, o tamanho dos túneis não interviu significativamente no método, pois a figura 34 não estabelece correlação significativa ($r \approx 0,30$) entre largura máxima de túnel e RMSD de reconstrução em harmônicos esféricos; logo, as dimensões dos túneis não influenciaram fortemente as distorções da reconstrução final nesta metodologia. Ademais, as distorções resultantes de fechamentos de túneis são apenas locais, afetando apenas a vizinhança de tais regiões. A técnica de fechamento de túneis de Benseddik et al.,¹⁰⁹ por exemplo, devido a etapa intermediária de voxelização seguida de preenchimento de *voxels* e posterior conversão em malha triangular, pode causar distorções em toda a superfície do objeto e alterações nos detalhes finos.

No contexto de atracamento de proteínas, contudo, esta metodologia exclui a possibilidade de penetração entre as superfícies, uma vez que é esperada apenas baixo grau de penetração. Caso túneis forem relevantes para o atracamento, estratégias alternativas de representação de superfícies podem ser adotadas. A estratégia de Mousa et al.,¹¹⁰ por exemplo, segmenta a malha original em regiões do tipo *star-like*, sem portanto a presença de túneis. Cada região é então independentemente representada em harmônicos esféricos.

3.5 Reconstrução e visualização de superfícies

Somando-se às sete estruturas cujos túneis não puderam ser tampados (seção 3.4), há ainda outras 55 da coleção *Protein Docking Benchmark* em que AHSP falhou em calcular o mapeamento esférico. No total, 62 estruturas não puderam ser representadas em harmônicos esféricos, o que corresponde a cerca de 7% do total de superfícies da

Figura 33 – Passos para se fechar um túnel: (a) detalhe de uma região de 3BX7_r_b com um túnel; (b) vértices no *loop* de túnel são destacados; (c) um buraco resta após remoção de vértices de (b); (d) túnel fechado.

Fonte: Elaborada pelo autor.

coleção. Esta taxa de falha é aceitável para uma prova de conceito da técnica. As demais 858 estruturas de *Protein Docking Benchmark* foram reconstruídas com $L_{max} = 50$, e a figura 35 indica a distribuição de seu RMSD. Todas as reconstruções apresentaram RMSD inferior a 1 Å, sendo cerca de 84% delas (777 estruturas) com RMSD inferior a 0,3 Å.

As reconstruções anteriores foram obtidas calculando-se as funções coordenadas X , Y e Z nos pontos (θ, ϕ) originais do mapeamento esférico, para que o RMSD fosse corretamente calculado entre cada ponto da superfície molecular e sua respectiva reconstrução na superfície projetada em harmônicos esféricos. Entretanto, o uso esperado do método consiste na restauração das funções originais apenas a partir dos coeficientes de harmônicos esféricos, sem conhecimento do mapeamento. Distribuição equidistante de pontos (θ, ϕ) de reconstrução pelo domínio esférico é importante para a correta restauração de informações da função projetada em harmônicos esféricos. A figura 36 mostra pontos calculados por três algoritmos de amostragem. O primeiro, regular (a), amostra o espaço bidimensional de coordenadas (θ, ϕ) em iguais passos angulares, resultando em acentuada concentração de pontos nas regiões circumpolares e escassa concentração no equador. Coordenadas (θ, ϕ) selecionadas aleatoriamente (b) tendem a exibir comportamento similar. A concentração

Figura 34 – Comparação entre largura máxima de túnel e RMSD de reconstruções com $L_{max} = 50$ de estruturas de *Protein Docking Benchmark*.

Fonte: Elaborada pelo autor.

Figura 35 – Distribuição de RMSD das reconstruções com $L_{max} = 50$ de estruturas de *Protein Docking Benchmark*.

Fonte: Elaborada pelo autor.

Figura 36 – Distribuição de pontos na superfície esférica usando três algoritmos de amostragem: (a) regular (b) seleção aleatória; (c) espiral de Fibonacci.

Fonte: Elaborada pelo autor.

de pontos em algumas regiões em detrimento de outras implica cálculos desnecessários realizados nestas regiões, pois, como os pontos são muito concentrados em relação às demais regiões, pouca informação é adicionada ao se calcular a reconstrução em cada um deles em comparação com calculá-la em um novo ponto onde a densidade é menor. Pontos selecionados pelo método da espiral de Fibonacci⁹⁴ (c) são mais homogeneamente espaçados e o cálculo da função de interesse em um novo ponto adiciona aproximadamente a mesma quantidade de informação da função representada, em qualquer região da esfera que seja calculada.

As estruturas da figura 37 (b)-(e), (g)-(j), (l)-(o), (q)-(t) e (v)-(y) foram reconstruídas seguindo o método da espiral a partir de superfícies originais (a, f, k, p, u) extraídas do *Protein Docking Benchmark*; para fins de visualização, uma malha triangular foi composta pelo método *Power crust*.⁹⁵ Não foram observadas distorções significativas e incompatíveis com a aproximação resultante da limitação do parâmetro l . O refinamento sucessivo de qualidade da representação, esperado conforme aumento de L_{max} , é prontamente percebido. Ainda na figura 37, as superfícies (a) e (u), que possuem quantidade mais numerosa de átomos e portanto mais detalhes finos, exigiram mais coeficientes de harmônicos esféricos para atingir o mesmo grau de fidelidade que a superfície (k), com menos átomos e portanto menos detalhes finos. Por este motivo, a estrutura da figura 37 (o) esboça o contorno de alguns átomos, ao passo que a estrutura da figura 37 (y) indica apenas detalhes mais nebulosos.

A tabela 5 caracteriza as cinco estruturas de exemplo em termos de quantidade de átomos, quantidade de vértices e triângulos da malha. O espaço de armazenamento necessário é da ordem de 10^4 e 10^3 KB. As reconstruções com ordem 40, por outro lado, requerem cada uma $3 \times (L_{max} + 1)^2 = 5043$ coeficientes de harmônicos esféricos, que

Figura 37 – Superfícies moleculares originais (a) 1DE4_l_u, (f) 1ZM4_r_u, (k) 3BX7_r_b, (p) 3G6D_r_u e (u) 4GAM_r_u, e suas reconstruções em harmônicos esféricos com (b, g, l, q, v) $L_{max} = 10$, (c, h, m, r, w) $L_{max} = 20$, (d, i, n, s, x) $L_{max} = 30$ e (e, j, o, t, y) $L_{max} = 40$.

Fonte: Elaborada pelo autor.

Tabela 5 – Detalhes de estruturas selecionadas

Nome	Átomos	Vértices	Triângulos	Espaço (KB)
1DE4_l_u	10112	224187	448370	10509
1ZM4_r_u	6419	130950	261896	6139
3BX7_r_b	1389	99179	198358	4650
3G6D_r_u	3234	174211	348418	8167
4GAM_r_u	17307	208226	416448	9761

Fonte: Elaborada pelo autor.

ocupam cerca de 40KB de espaço de armazenamento, uma diferença de ao menos duas ordens de grandeza. Esta economia não é gratuita, contudo, e só é possível após as etapas de cálculo de superfície, mapeamento esférico e cálculo de coeficientes. Este custo inicial se justifica em *docking* de *ensemble* de proteínas pela facilidade de refinamento sucessivo, que permite eliminar orientações desfavoráveis nos estágios iniciais do procedimento, e pela decomposição espectral da seção 3.8.

3.6 Cálculo de curvaturas da superfície

O cálculo das curvaturas gaussiana e média seguiu as fórmulas propostas por Duncan e Olson,⁴⁶ resolvidas por diferenciação numérica a partir dos coeficientes de harmônicos esféricos, e aplicadas nas superfícies reconstruídas de 1R2M da figura 38 (c) e (e) para curvatura gaussiana e da figura 38 (d) e (f) para curvatura média, com $L_{max} = 20$ e $L_{max} = 50$, respectivamente. As curvaturas para a superfície original das figuras 38 (a) e (b) foram calculadas por CGAL a partir dos pontos da malha original. A escala de cor de cada figura foi normalizada pelo maior valor absoluto de cada situação para identificar o sinal (positivo ou negativo) da curvatura. Com isto, uma mesma cor pode indicar valores absolutos de curvatura diferentes em outras figuras, porém sem prejuízo ao seu sinal. A escolha de cores foi arbitrária e apenas para fins de visualização; internamente, o programa lidou com valores numéricos de curvatura ao invés de cores. A curvatura gaussiana identifica corretamente regiões de domo em azul e regiões de sela em vermelho, ao passo que a curvatura média aponta cavidades em vermelho e protuberâncias em azul, confirmando que curvaturas de superfície podem ser calculadas sobre os coeficientes de harmônicos esféricos apenas. Em uma abordagem de atracamento de proteínas, isto elimina a necessidade de reconstrução de superfícies para aferição de complementaridade geométrica entre as moléculas.

Para fins de comparação, CGAL foi usado para calcular as curvaturas gaussiana e média a partir dos pontos reconstruídos de 1R2M. A figura 39 (a) estabelece elevada correspondência geral ($r \approx 0,94$) entre os resultados de curvatura gaussiana calculados

Figura 38 – Curvaturas gaussianas (a, c, e) e média (b, d, f) na superfície de 1R2M: (a, b) superfície original; (b, c) reconstruções com $L_{max} = 20$; (e, f) reconstruções com $L_{max} = 50$.

Fonte: Elaborada pelo autor.

pelos dois métodos em reconstrução com $L_{max} = 20$, no entanto com alguns valores com valor absoluto incorreto, ainda que com sinal adequado. O mesmo se observou para curvatura média ($r \approx 0,98$). Já o aumento de funções de base na reconstrução com $L_{max} = 50$ causou ligeiro recuo da correlação de curvatura média ($r \approx 0,89$) e recuo mais acentuado na correlação dos valores de curvatura gaussiana ($r \approx 0,70$, figura 39 (b)). Ainda que a maioria dos resultados tenham sido adequados, cabe refinamento na técnica de diferenciação numérica empregada.

3.7 Cálculo de energia potencial eletrostática

A superfície eletrostática, utilizada na determinação de afinidade de interação entre moléculas, foi calculada através da biblioteca BALL para 1R2M, como ilustrado pela figura 40 (a), e sua energia potencial foi reconstruída em harmônicos esféricos. As figuras 40 (b-d) mostram as reconstruções com $L_{max} = 20, 30$ e 50 , respectivamente, em que tanto as funções coordenadas que definem a geometria da superfície quanto a função de energia potencial foram representadas em harmônicos esféricos. Assim como as

(a)

(b)

Figura 39 – Comparação entre curvatura gaussiana calculada por CGAL e calculada por coeficientes de harmônicos esféricos para a superfície reconstruída de 1R2M com (a) $L_{max} = 20$ e (b) $L_{max} = 50$.

Fonte: Elaborada pelo autor.

Figura 40 – Distribuição de energia potencial eletrostática pela superfície de 1R2M: (a) superfície original; (b) reconstrução com $L_{max} = 20$; (c) reconstrução com $L_{max} = 30$; (d) reconstrução com $L_{max} = 50$.

Fonte: Elaborada pelo autor.

superfícies geométricas calculadas nos tópicos anteriores são representações compatíveis com as respectivas superfícies originais e com a limitação do parâmetro l , as superfícies eletrostáticas reconstruídas concordam com a superfície eletrostática original, e o aumento de L_{max} melhora a correspondência entre elas. As correlações obtidas foram de 0,58 ($L_{max} = 20$), 0,70 ($L_{max} = 30$) e 0,87 ($L_{max} = 50$), com comportamento semelhante às das reconstruções das funções coordenadas X , Y e Z . Uma vez a energia potencial eletrostática calculada, não é necessário mais recalculá-la com a invocação de BALL, apenas reconstruí-la nas regiões desejadas a partir dos coeficientes de harmônicos esféricos conforme o mesmo procedimento de restauração de geometria de superfície.

3.8 Decomposição espectral de superfícies de um *ensemble* de conformações

A decomposição espectral da seção 2.10 foi aplicada em 232 conformações do *ensemble* de identificação PDB 2LJ5 de ubiquitina,²⁴ uma proteína rígida globular que apresenta apenas a cauda N-terminal como região flexível. As demais 68 conformações, cerca de 23% do total, causaram erros no estágio de mapeamento esférico, novamente associados à presença de túneis. Segundo o critério de distância, a estrutura de número 271 foi selecionada como referência. Após o cálculo de coeficientes de harmônicos esféricos para todas as estruturas, distância máxima e MAD foram calculados para todos os coeficientes relativos a cada eixo de coordenadas e com mesmo grau e mesma ordem, ou seja, mesmos valores de l e mesmos valores de m , para todas as 232 superfícies. A figura 41 mostra o RMSD médio e correspondente desvio padrão obtido após reconstrução das superfícies de

2LJ5 com $L_{max} = 40$ e diferentes quantidades de coeficientes da estrutura de referência para as duas métricas de variabilidade examinadas, distância máxima na figura 41 (a) e MAD¹⁰¹ na figura 41 (b). Para cada porcentagem p , $(100 - p)\%$ dos coeficientes se originaram dos coeficientes mais variáveis de cada estrutura e os $p\%$ restantes foram tomados da estrutura de referência. Como esperado, o uso mais numeroso de coeficientes da estrutura de referência foi acompanhado de aumento de RMSD da respectiva reconstrução. Em todos os valores de p as reconstruções a partir de variabilidade determinada por MAD foram em média mais similares às superfícies originais do que as reconstruções com coeficientes selecionados por distância máxima, sugerindo portanto MAD como uma métrica de variabilidade mais adequada do que distância máxima.

O espalhamento de valores dos coeficientes de harmônicos esféricos é ilustrado na figura 42 para os de maior variação e na figura 43 para os de menor variação, em ambas em figuras em relação a X, Y e Z. A largura das barras dos histogramas foi selecionada automaticamente.¹¹¹ O intervalo de valores dos coeficientes de menor variação, indicados nas figuras 43 (a-c), foram idênticos nas duas métricas. No entanto, os intervalos de valores das figuras 42 (a), 42 (c) e 42 (e) (à esquerda), são mais amplos do que os intervalos de valores das figuras 42 (b), 42 (d) e 42 (f) (à direita), e confirmam captura mais adequada de variabilidade dos coeficientes pela métrica MAD.

Nas reconstruções com MAD da figura 41 (b), as três situações resultaram em aumento de RMSD sem saltos excessivos conforme aumento da porcentagem de coeficientes p . No entanto, as reconstruções após filtragem por *wavelets* foram em média mais similares às superfícies originais do que sem filtragem, com ligeira vantagem de Daubechies para $p \geq 40$. Com exceção de $p = 10$ e $p = 90$, a diferença relativa entre Haar e Daubechies é diminuta. No *ensemble* utilizado houve benefício na filtragem de MAD antes de ordenação e seleção de coeficientes segundo sua variabilidade.

O efeito de variação da quantidade de coeficientes em uma única superfície é demonstrado pela figura 44 para reconstruções, após medição de variabilidade com MAD e posterior filtragem com *wavelet* Daubechies, da estrutura de número 0 do *ensemble*, e pela figura 45 para reconstruções da estrutura de número 71. O aumento de p torna as estruturas menos parecidas com suas reconstruções usando apenas coeficientes próprios, em (a), porém sem variações visuais bruscas entre os incrementos testados de p , o que é compatível com os incrementos de RMSD observados na figura 41 (b). A superfície de (g), que utiliza 50% de coeficientes da estrutura de referência, mantém o aspecto geral da superfície original em (a). Já a superfície de (j), ainda que compartilhe 90% de seus coeficientes com a superfície de referência em (k), é altamente distorcida. É notável que nas reconstruções do modelo 0, da figura 44, a parte superior, referente à cauda N-terminal da proteína, mantém-se inclinada para a esquerda como em (a), mesmo com $p = 80$. Tendências similares de redução de qualidade de reconstrução com aumento de p

(a)

(b)

Figura 41 – RMSD de reconstruções com $L_{max} = 40$ de 2LJ5 com número variado de coeficientes da estrutura de referência usando (a) distância máxima e (b) MAD.

Fonte: Elaborada pelo autor.

Figura 42 – Distribuição de valores para o coeficiente de maior variação segundo métrica MAD (a, c, e) e segundo métrica de máxima distância (b, d, f) em X (a, b), Y (c, d) e Z (e, f).

Fonte: Elaborada pelo autor.

(a)

(b)

(c)

Figura 43 – Distribuição de valores para o coeficiente de menor variação segundo métrica MAD em (a) X , (b) Y e (c) Z . As distribuições segundo métrica de máxima distância foram iguais às segundo MAD.

Fonte: Elaborada pelo autor.

Figura 44 – Reconstruções de superfícies de 2LJ5 para $L_{max} = 40$ com seleção de coeficientes segundo MAD após filtragem por Daubechies: (a) estrutura 0; (b-j) reconstruções da estrutura 0 usando (b) 10%, (c) 20%, (d) 30%, (e) 40%, (f) 50%, (g) 60%, (h) 70%, (i) 80%, e (j) 90% de coeficientes da estrutura de referência; (k) estrutura de referência (271).

Fonte: Elaborada pelo autor.

também foram observadas para menores quantidades de bandas de harmônicos esféricos, conforme figura 46 e $L_{max} = 10, 20$ e 30 . Uma estratégia de *docking* de proteínas poderia realizar reconstruções grosseiras das superfícies nas primeiras iterações do método e se aproveitar da baixa quantidade de detalhes finos para utilizar uma quantidade elevada de coeficientes comuns a todas as estruturas do *ensemble*, combinando, assim, baixo L_{max} com elevada porcentagem p . O refinamento das superfícies com aumento de L_{max} após eliminação de orientações desfavoráveis seria acompanhado de redução de p . Ao se usar por exemplo 30% de coeficientes da estrutura de referência, uma estrutura, a de referência, é representada por 100% de seus coeficientes, enquanto todas as demais são representadas por apenas 70% dos coeficientes, resultando em uma economia de 30% de memória e de processamento e, neste exemplo específico mantendo fidelidade adequada às superfícies originais. Uma abordagem automática para escolha de p para um dado *ensemble* evitaria uma simples porém ineficiente busca exaustiva pelo valor de p que resulte em reconstrução mais adequada, e é proposta como um desenvolvimento futuro.

As reconstruções com filtragem de métrica MAD resultaram em média em maior fidelidade. Ao contrário de sinais temporais, os coeficientes de harmônicos esféricos não possuem, contudo, uma ordenação que corresponda à ordem utilizada anteriormente. A figura 47 mostra o resultado da filtragem por Daubechies da métrica MAD seguindo a ordem de listagem de coeficientes $c_l^m = c_i$, $i = l(l + 1) + m$,⁴⁵ e indica que diversos coeficientes tiveram seu valor de MAD fortemente reduzido, o que introduz no conjunto de

Figura 45 – Reconstruções de superfícies de 2LJ5 para $L_{max} = 40$ com seleção de coeficientes segundo MAD após filtragem por Daubechies: (a) estrutura 71; (b-j) reconstruções da estrutura 71 usando (b) 10%, (c) 20%, (d) 30%, (e) 40%, (f) 50%, (g) 60%, (h) 70%, (i) 80%, e (j) 90% de coeficientes da estrutura de referência; (k) estrutura de referência (271).

Fonte: Elaborada pelo autor.

(a)

Figura 46 – RMSD de reconstruções com MAD e filtragem Daubechies de 2LJ5 com número variado de coeficientes da estrutura de referência usando $L_{max} = 10, 20$ e 30 .

Fonte: Elaborada pelo autor.

dados uma nova ordenação de valores segundo MAD. Ao contrário do que se esperava, por contrariar a ordenação segundo MAD, esta nova ordem melhorou a qualidade das reconstruções, conforme menores RMSD médios da figura 41 (b). Se os coeficientes tivessem sido reunidos em ordem distinta, a curva de MAD filtrada teria sido diferente, com potencial influência na qualidade das reconstruções. Isto sugere que a hierarquia de contribuição dos coeficientes na representação final da função projetada em harmônicos esféricos não segue a mesma ordenação dada por MAD. Para investigar tal hipótese, foram feitas reconstruções da estrutura de número 271 de 2LJ5, escolhida por ter sido a referência nos parágrafos anteriores, utilizando apenas um único coeficiente. Para ordem de reconstrução $L_{max} = 40$, a reconstrução i utilizou grau l e ordem m tal que $i = l(l+1) + m$, totalizando $(L_{max} + 1)^2 = 1681$ reconstruções. O RMSD de reconstrução resultante, em ordem decrescente, foi considerado como métrica de variabilidade dos coeficientes e utilizado na decomposição espectral das superfícies do *ensemble* 2LJ5 em substituição à distância máxima e a MAD. Não foi realizada filtragem por *wavelet*. A figura 48 mostra que as reconstruções resultantes desta ordenação de coeficientes foram em média mais similares às superfícies originais, isto é, com menores valores médios de RMSD, do que a melhor técnica dos parágrafos anteriores, que organizou os coeficientes de harmônicos esféricos segundo MAD após filtragem por *wavelet* Daubechies. Ademais, o desvio padrão das novas reconstruções também foi inferior em relação à reconstrução anterior, indicando que menos superfícies foram prejudicadas pelo uso de coeficientes alheios. Ao se analisar as superfícies resultantes nas figuras 49 e 50, nota-se que o aspecto geral da superfície, incluindo a maioria das protuberâncias, é melhor representado por esta técnica em comparação com MAD e Daubechies, especialmente em elevadas porcentagens. Contudo, os detalhes mais finos, como contorno dos átomos, são melhor capturados por MAD e Daubechies. As distorções obtidas foram distintas entre os métodos, indicando que a combinação de coeficientes de superfícies diferentes precisa ser feita de forma cuidadosa.

Decomposição espectral é um método promissor no contexto de *docking* de proteínas, mas exige investigação mais extensa, por exemplo utilizando medidas diferentes de variabilidade de acordo com a porcentagem de coeficientes a usar da estrutura de referência. Este desenvolvimento é sugerido como um estudo futuro.

3.9 Conclusões

Neste capítulo foi realizada uma análise da influência dos diversos componentes da representação parametrizada em harmônicos esféricos com o objetivo de obtenção de elevada qualidade de reconstrução no contexto de atracamento de proteínas. Foi identificada como o principal fator determinante de qualidade de reconstruções a distribuição de amostras da função representada pela superfície esférica. Distribuições acentuadamente irregulares produziram distorções nos resultados. Apenas os mapeamentos com maior regularidade

Figura 47 – Curvas de MAD a partir de valores originais e após filtragem por *wavelet* de Daubechies.

Fonte: Elaborada pelo autor.

resultaram em reconstruções com qualidade adequada e compatível com o truncamento representado pela limitação do grau l . No caso de malhas triangulares, esta regularidade se observa, também, pela distorção de área de seus triângulos constituintes. Portanto, é necessário um algoritmo de mapeamento que se disponha a reduzir a distorção de área.

A obtenção de um mapa com distribuição homogênea de pontos pela superfície esférica é, contudo, extremamente difícil. O método de transformação de superfície de [Duncan e Olson⁴⁶](#) é impreciso, pois não realiza um mapeamento esférico, mas apenas uma deformação que elimina a maioria das regiões *star-like*, no entanto sem efetivamente garantir a unicidade da função em todo o domínio esférico. Como consequência, não há cuidado com distribuição regular de pontos após a deformação. O mapeamento CALD resultou em distorções e reconstruções insatisfatórias motivadas por baixa regularidade na distribuição de pontos. Não obstante, é usado em reconstruções com baixo nível de detalhes, como, por exemplo, ordem de reconstrução igual a 15 de grãos de areia pouco detalhados¹¹² e representação de densidade de ligantes em torno de uma proteína

(a)

Figura 48 – RMSD de reconstruções de 2LJ5 com número variado de coeficientes da estrutura de referência usando $L_{max} = 40$ e (a) seleção de coeficientes por MAD filtrado por Daubechies e (b) seleção de coeficientes segundo influência no RMSD de reconstruções da estrutura de referência.

Fonte: Elaborada pelo autor.

Figura 49 – Reconstruções de superfícies de 2LJ5 para $L_{max} = 40$ com ordenação de coeficientes segundo influência no RMSD: (a) estrutura 0; (b-j) reconstruções da estrutura 0 usando (b) 10%, (c) 20%, (d) 30%, (e) 40%, (f) 50%, (g) 60%, (h) 70%, (i) 80%, e (j) 90% de coeficientes da estrutura de referência; (k) estrutura de referência (271).

Fonte: Elaborada pelo autor.

Figura 50 – Reconstruções de superfícies de 2LJ5 para $L_{max} = 40$ com ordenação de coeficientes segundo influência no RMSD: (a) estrutura 71; (b-j) reconstruções da estrutura 0 usando (b) 10%, (c) 20%, (d) 30%, (e) 40%, (f) 50%, (g) 60%, (h) 70%, (i) 80%, e (j) 90% de coeficientes da estrutura de referência; (k) estrutura de referência (271).

Fonte: Elaborada pelo autor.

com ordem de reconstrução igual a 5.⁴⁴ ConformalizedMCF, ARAP, BLD e AHSP são desenvolvimentos mais recentes que atingiram progressiva melhora de regularidade de distribuição de vértices. Ainda que BLD tenha permitido reconstruções de qualidade superior nos dois únicos exemplos testados, o emprego de AHSP foi mais conveniente e resultou em mapeamentos com regularidade adequada. Diversos trabalhos têm lidado com o problema de mapeamento esférico, e é possível que novos surjam e forneçam ainda menores distorções.

A distribuição de vértices na malha triangular que determina a superfície molecular também é um fator de influência na qualidade dos resultados. Distribuições mais regulares colaboram para menores distorções de área no mapeamento e consequente distribuição mais homogênea de pontos na superfície esférica. A superfície calculada por CGAL é adequada para visualização, mas não para representação de superfícies pelas técnicas estudadas, pois apresenta densidade de pontos que varia conforme o grau de detalhamento. EDTSurf varia muito menos a densidade de pontos, resultando em malha triangular com regularidade apropriada para este projeto. Além disso, permite escolha do tipo de superfície.

O cálculo dos coeficientes de harmônicos esféricos por integração numérica exige algum método de interpolação para avaliar as funções em pontos onde seus valores são desconhecidos. Por outro lado, a minimização por quadrados mínimos do método IRF concentra-se em solucionar as equações para os pontos onde a função é conhecida, e nos testes apresentou resultados satisfatórios.

A principal limitação da metodologia original de representação paramétrica de superfícies tridimensionais em harmônicos esféricos⁴⁶ é a presença de túneis, que inibem a deformação da superfície em um formato esférico. Neste trabalho foi mostrado que túneis podem ser cobertos, tão logo quanto puderem ser ignorados, estendendo, assim, a validade da metodologia para superfícies de topologias arbitrárias. Túneis muito grandes causam maiores distorções, contudo. Ainda que no conjunto de dados *Protein Docking Benchmark* a taxa de falhas tenha sido discreta, no conjunto menor dado pelo *ensemble* de 2LJ5 a taxa de falhas foi significativa, indicando oportunidade de melhoria tanto no fechamento de túneis quanto no mapeamento esférico. Os resultados obtidos com a nova combinação de técnicas indicam alta qualidade de reprodução do relevo das superfícies, e seu uso em superfícies de variados formatos sugere aplicabilidade em relevos arbitrários.

Para algumas aplicações, baixas ordens de reconstrução são suficientes.^{112,113} Para proteínas, no entanto, bandas mais numerosas são necessárias para se atingir detalhamento fino. A maioria das reconstruções deste capítulo foi realizada com $L_{max} = 50$, resultando em $3 \times 2601 = 7803$ coeficientes e apresentando elevada correspondência com as superfícies originais, ainda que detalhes mais finos estejam ausentes. Reconstruções com ordens de reconstrução elevadas, superiores a 100, não foram testadas. Em alguns casos reconstrução resultou em explosão de valores já a partir de $L_{max} = 78$. Esta representação é compacta, pois o volume de dados para ordem reconstrução igual a 50 é uma ordem de grandeza inferior à da quantidade de vértices na superfície da menor estrutura de *Protein Docking Benchmark*, que possui 22259 pontos, e duas ordens de grandeza inferior aos 276514 vértices da superfície da maior estrutura desta coleção. A combinação de refinamento sucessivo de qualidade com decomposição espectral reduz grandemente a quantidade de dados a manipular após obtenção de coeficientes de harmônicos esféricos e permite reconstruir superfícies com grau ajustável de detalhamento. Em uma estratégia de atracamento de proteínas, isto significa permitir detecção, em baixo nível de detalhes, de orientações mais favoráveis das moléculas e eliminação de orientações desfavoráveis nas etapas seguintes de busca com superfícies mais detalhadas. Uma vez a molécula representada deste modo, as etapas posteriores de *docking* são conhecidas.⁵⁵ Mesmo representações em harmônicos esféricos limitadas a superfícies *star-like* foram usadas para *docking*,^{32,48} o que torna a representação paramétrica ainda mais atraente para este contexto.

Há interesse da representação em outras áreas além de *docking* de proteínas, como, por exemplo, reconstrução de grãos de areia para estudos de engenharia civil¹¹² e comparação de superfícies de imagens médicas e detecção de mudanças associadas com doença de Alzheimer.¹¹³ Melhorias na representação paramétrica podem beneficiar outras áreas de conhecimento.

4 REPRESENTAÇÃO DE SUPERFÍCIES MOLECULARES EM ZERNIKE 3D COM MAPEAMENTO NA BOLA UNITÁRIA

A etapa de mapeamento esférico da representação paramétrica em harmônicos esféricos dos capítulos 2 e 3 introduz complexidade e distorções causadas pela distribuição irregular de vértices da malha original na superfície esférica. Os túneis das superfícies representam dificuldade adicional. Este capítulo indica a metodologia de análise de representações existentes de superfícies em polinômios tridimensionais de Zernike, que não exigem a etapa de mapeamento esférico e portanto são diretamente aplicáveis a superfícies com qualquer topologia, e propõe duas novas representações. As discussões se iniciam na seção 4.1 com descrição dos procedimentos para avaliação dos polinômios de Zernike 3D. A representação tradicional com estes polinômios para objetos volumétricos binários tem seus métodos apresentados na seção 4.2, que trata da conversão de malha tridimensional em uma representação volumétrica, e na seção 4.3, que trata de cálculo de coeficientes e reconstrução de funções binárias. A reconstrução de funções gerais é realizada seguindo a seção 4.4. Para melhorar a qualidade das reconstruções, duas novas formas de se calcular os momentos de Zernike são propostos com base na reordenação dos valores da função projetada, primeiramente na seção 4.5 para reordenação axial de valores, e em seguida na seção 4.6, para reordenação espiral.

4.1 Avaliação de integração

Como complemento à seção 2.2, uma investigação foi feita acerca da normalidade e ortogonalidade de polinômios tridimensionais de Zernike. Sua relação de ortogonalidade é dada pela equação 1.16,¹¹⁴ indicando que a integral de polinômios com mesmos índices devem ser igual a $\frac{4\pi}{3}$, e igual a zero caso os índices sejam diferentes. Contudo, assim como na seção 2.6, o cálculo da equação 1.16 diretamente a partir da definição analítica de polinômios tridimensionais de Zernike incorre em erros numéricos no cálculo dos polinômios de Legendre para valores elevados de n . Para contorná-los, novamente utilizaram-se algoritmos adequados^{51,87} combinados com Monte Carlo para calcular a equação 1.16 por aproximação numérica. Os pontos de integração foram selecionados pelo algoritmo *ball_grid*.¹¹⁵ Procedimento similar foi usado para se verificar a normalidade das funções radiais, conforme equação 1.17, neste caso utilizando pontos de amostragem selecionados pelo algoritmo *sphere_grid*.¹¹⁶

Duas definições de funções radial são conhecidas, a das equações 4.1 e 4.2,^{60,114} e a das equações 4.3 e 4.4,⁶¹ ambas para $(n - l)$ par, ou zero caso contrário. As duas definições

foram utilizadas na solução da integral de ortonormalidade.

$$R_{nl}(r) = r^l \sum_{v=0}^k q_{kl}^v r^{2v}, \quad k = \frac{n-l}{2} \quad (4.1)$$

$$q_{kl}^v = \frac{(-1)^k}{2^{2k}} \sqrt{\frac{2l+4k+3}{3}} \binom{2k}{k} (-1)^v \frac{\binom{k}{v} \binom{2(k+l+v)+1}{2k}}{\binom{k+l+v}{k}} \quad (4.2)$$

$$R_{nl}(r) = \sum_{k=0}^{(n-l)/2} N_{nlk} r^{n-2k} \quad (4.3)$$

$$N_{nlk} = (-1)^k 2^{l-n} \sqrt{2n+3} \times \frac{(2n-2k+1)! [(1/2)(n+l)-k]!}{[(1/2)(n-l)-k]! (n+l-2k+1)! (n-k)! k!} \quad (4.4)$$

Métodos adaptativos de integração aplicam regras de quadratura em subdomínios do domínio original de integração até que convergência seja atingida,^{117,118} desta forma lidando melhor com reduzida suavidade da função de integração. A biblioteca *cubature*¹¹⁹ implementa este mecanismo e foi usada para avaliar numericamente a equação 1.16 através de sua rotina `hcubature` e, com isso, aumentar a confiança nos testes de integração. Uma desvantagem de tal abordagem é a elevada demanda imposta a recursos de memória, que aumenta conforme o domínio de integração é sucessivamente dividido. Para permitir o uso da biblioteca com a memória disponível no computador, o domínio inicial (a bola unitária) foi dividido em fatias de igual tamanho, cada uma delas usada como entrada para um processo que invoca `hcubature`. Ao término das execuções, todas as integrações em subdomínios foram somadas para produzir o resultado final.

4.2 Voxelização

A representação de superfícies em Zernike 3D é centrada no cálculo dos momentos Ω_{nl}^m através da equação 4.5. Uma função $f(\mathbf{x}) : \mathbb{R}^3 \rightarrow \mathbb{R}$ pode ser representada aproximadamente por uma grade volumétrica, em que o volume total é dividido em regiões cúbicas de igual tamanho chamadas de voxel (do inglês *volumetric element*, ou elemento volumétrico). Cada voxel possui uma coordenada tridimensional, dada pela posição de seu centroide $\mathbf{x}$, e é associado ao valor da função $f(\mathbf{x})$. Caso $f(\mathbf{x})$ seja uma função binária, seu conjunto imagem pode ser interpretado como definindo um objeto tridimensional. A voxels que não pertencem ao objeto é então atribuído o valor 0, ao passo que voxels dele constituintes podem indicar um valor constante, como, por exemplo, 1. A função $f(\mathbf{x})$ também pode definir qualquer outra propriedade no espaço tridimensional. Sua descrição em voxels exige de sua reconstrução em Zernike 3D apenas um único conjunto de momentos Ω_{nl}^m .

$$\Omega_{nl}^m = \frac{3}{4\pi} \int_{|\mathbf{x}| \leq 1} f(\mathbf{x}) \overline{Z_{nl}^m(\mathbf{x})} d\mathbf{x} \quad (4.5)$$

Uma superfície molecular descrita por malha triangular obtida de EDTSurf pode ser trivialmente voxelizada. Uma vez escolhida a resolução N_{grid} da grade volumétrica com $(N_{grid})^3$ voxels, divide-se as coordenadas de cada centroide de triângulo pela largura de um voxel para determinar a qual voxel tal triângulo pertence. Voxels populadas recebem o valor constante 1 ou o valor de $f(\mathbf{x})$, de acordo com a interpretação de f , ao passo que voxels vazios recebem o valor 0. Voxels no interior da molécula, que podem ser necessários para sua correta representação em Zernike 3D, recebem o valor 1 (ou o valor de $f(\mathbf{x})$); no entanto, exigem trabalho mais elaborado. Cada fileira de voxels é percorrida na direção paralela ao eixo Z. Ao encontrar um voxel preenchido, consulta-se o vetor normal, obtido da superfície em malha. Vetores normais apontando em sentido contrário ao da varredura indicam que se está entrando no volume da molécula, ao passo que voxels apontando no mesmo sentido da varredura indicam que se está saindo do volume. Cada par de voxels em que o sentido do vetor normal do primeiro é contrário e do segundo é igual ao da varredura delimita uma área interior à molécula. Por fim, os voxels são armazenados em um *array*.

As descrições em malha triangular de superfícies moleculares foram calculadas a partir de arquivos PDB pelo programa EDTSurf.⁷³

4.3 Reconstrução de objetos volumétricos binários

Novotni e Klein¹¹⁴ propuseram um algoritmo prático para cálculo dos momentos de Zernike. A equação 4.6 determina os polinômios 3D de Zernike em coordenadas cartesianas e a equação 4.7 determina os momentos de Zernike para uma função geral $f(\mathbf{x})$, com $r = 2(v + \alpha) + u$, $s = 2(\mu - v + \beta) + m - u$ e $t = 2(v - \alpha - \beta - \mu) + l - m$. Com isso, os coeficientes χ_{nlm}^{rst} não dependem das coordenadas (x, y, z) dos pontos e podem ter seus cálculos custosos, mostrados na equação 4.8, realizados previamente e reutilizados para qualquer função f objeto de reconstrução. Os fatores de normalização c_l^m são dados pela equação 4.9.

$$Z_{nl}^m(\mathbf{x}) = \sum_{r+s+t \leq n} \chi_{nlm}^{rst} x^r y^s z^t \quad (4.6)$$

$$\Omega_{nl}^m = \frac{3}{4\pi} \sum_{r+s+t \leq n} \chi_{nlm}^{rst} \int_{|\mathbf{x}| \leq 1} f(\mathbf{x}) x^r y^s z^t d\mathbf{x} \quad (4.7)$$

$$\begin{aligned} \chi_{nlm}^{rst} = & c_l^m 2^{-m} \sum_{v=0}^k q_{kl}^v \sum_{\alpha=0}^v \binom{v}{\alpha} \cdot \sum_{\beta=0}^{v-\alpha} \binom{v-\alpha}{\beta} \sum_{u=0}^m (-1)^{m-u} \binom{m}{u} i^u \\ & \cdot \sum_{\mu=0}^{\lfloor \frac{l-m}{2} \rfloor} (-1)^\mu 2^{-2\mu} \binom{l}{\mu} \binom{l-\mu}{m+\mu} \sum_{v=0}^{\mu} \binom{\mu}{v} \end{aligned} \quad (4.8)$$

$$c_l^m = \frac{\sqrt{(2l+1)(l+m)!(l-m)!}}{l!} \quad (4.9)$$

O método da equação 4.7 tira vantagem da regularidade da grade volumétrica e da simplicidade dos monômios x^r , y^s e z^t para obter analiticamente suas antiderivadas e, com isso, aproximar a integral de cálculo dos momentos por uma soma discreta sobre os voxels. Para uma grade volumétrica com $(N_{grid})^3$ voxels, os momentos geométricos G_{rst} (equação 4.10) podem ser substituídos por M_{rst} (equação 4.11), permitindo calcular os momentos de Zernike 3D pela equação 4.12. A função original é reconstruída com ordem N_{max} pela equação 4.13; os valores reconstruídos f_{ZK} são comparados a um limiar para se remover ruído.⁶¹

$$G_{rst} = \int_{|\mathbf{x}| \leq 1} f(\mathbf{x}) x^r y^s z^t d\mathbf{x} \quad (4.10)$$

$$M_{rst} = \sum_{i=0}^{N_{grid}-1} \frac{x_{i+1}^{r+1} - x_i^{r+1}}{r+1} \cdot \sum_{j=0}^{N_{grid}-1} \frac{y_{j+1}^{s+1} - y_j^{s+1}}{s+1} \cdot \sum_{k=0}^{N_{grid}-1} \frac{z_{k+1}^{t+1} - z_k^{t+1}}{t+1} \quad (4.11)$$

$$\Omega_{nl}^m = \frac{3}{4\pi} \sum_{r+s+t \leq n} \overline{\chi_{nlm}^{rst}} M_{rst} \quad (4.12)$$

$$f(\mathbf{x}) \approx f_{ZK}(\mathbf{x}) = \sum_{n=0}^{N_{max}} \sum_{l=0}^n \sum_{m=-l}^l \Omega_{nl}^m Z_{nl}^m(\mathbf{x}) \quad (4.13)$$

Este algoritmo foi utilizado para representar superfícies moleculares voxelizadas pelos passos da seção 4.2, através da implementação de referência,¹¹⁴ da implementação de SASTBX,¹²⁰ e de implementação independente feita pelo estudante.

4.4 Reconstrução de objetos volumétricos definidos por funções gerais

De modo similar à representação paramétrica em harmônicos esféricos, a malha triangular de superfícies moleculares calculadas por EDTSurf foi representada em Zernike 3D pelo algoritmo da seção 4.3, projetando-se separadamente as funções X , Y e Z e resultando em três conjuntos de momentos, $\Omega_{x,nl}^m$, $\Omega_{y,nl}^m$ e $\Omega_{z,nl}^m$. A filtragem de valores por limiar foi, naturalmente, omitida. Além de SASTBX e da implementação de Zernike 3D do estudante, foi avaliado também o programa de Nisius e Gohlke,¹²¹ que compartilha a maioria de seu código com a implementação de referência.¹¹⁴

4.5 Mapeamento na bola unitária com reordenação axial

Uma vez que reconstruções usando polinômios tridimensionais de Zernike devem considerar todo o espaço no interior da esfera, isto é, a bola, uma abordagem distinta

foi planejada. Ao invés de calcular a equação 4.5 usando as coordenadas originais dos vértices, a malha triangular foi colapsada em uma bola, ou seja, deformada em um formato esférico, mas desta vez também preenchendo seu espaço interior. Isto requer que os vértices sejam ordenados de um modo específico para que a função projetada, seja X , Y ou Z , seja não decrescente. Assim como no capítulo 2, EDTSurf foi usado para criar a malha triangular inicial dadas as coordenadas atômicas em um arquivo PDB. Sucintamente, isto é feito posicionando-se os átomos em uma grade tridimensional e em seguida aplicando-se o algoritmo *marching cubes*⁷⁴ para converter um conjunto de pontos em uma malha conectada. Internamente, EDTSurf visita os vértices de acordo com três *loops*, ou laços de repetição, aninhados em seu código-fonte. O *loop* mais externo percorre células da grade na direção do eixo X ; o *loop* do meio percorre células na direção do eixo Y ; finalmente, o *loop* interno percorre as células na direção do eixo Z . Como a função X é representada no *loop* externo, seus valores serão sempre não decrescentes. Para se projetar a função Y , o segundo e o primeiro laços foram substituídos um pelo outro, de modo a posicionar o laço referente à direção Y na posição mais externa. Para se projetar a função Z , o laço interno foi movido para a posição mais externa. Três rotinas independentes ficaram estabelecidas, portanto, cada uma percorrendo pontos de um eixo coordenado no *loop* mais externo. A bola unitária, e portanto seu interior também, foi considerada como uma lista linear e preenchida com amostras da função de interesse na nova ordem mencionada. A resolução N_{grid} da grade de projeção foi escolhida automaticamente de forma a comportar todos os valores da função. Deste modo, quase todo o espaço de amostragem contém informação relevante sobre a função. O espaço vazio que não contribui para o valor desejado dos momentos, devido à diferença entre o total de *voxels* da grade e o total de valores da função, é sensivelmente diminuído. Há uma ressalva, contudo: para se construir o objeto corretamente, é preciso reordenar os valores reconstruídos de Y e Z para refletir a mesma ordem de X e assim formar um ponto válido (x, y, z) . Isto foi facilmente resolvido com a introdução de uma mapa associando X a Y e outro mapa associando X a Z em termos de posição na lista linear que contém os dados da bola unitária. Tais mapas não devem ser confundidos com os mapas esféricos do capítulo 2, contudo. O objeto reconstruído desta forma pode então aproximar a geometria do objeto original.

4.6 Mapeamento na bola unitária com reordenação espiral

O método da seção 4.5 introduz dados de mapeamento que são alheios à representação em Zernike 3D. Para eliminá-los sem se perder a associação entre as coordenadas x_i , y_i e z_i do ponto $\mathbf{x}$, e ainda assim permitir mapeamento na bola unitária, uma reordenação de $f(\mathbf{x})$ baseada em espiral retangular foi criada. A figura 51 mostra duas grades retangulares de tamanho 5×5 cujas células foram visitadas por espirais retangulares, de movimento para o interior em (a) e de movimento para o exterior em (b). Os números indicam a ordem de visitação a partir da passo inicial 0. Após voxelização individual das funções

0	1	2	3	4	20	21	22	23	24
15	16	17	18	5	19	6	7	8	9
14	23	24	19	6	18	5	0	1	10
13	22	21	20	7	17	4	3	2	11
12	11	10	9	8	16	15	14	13	12
(a)					(b)				

Figura 51 – Espirais retangulares de sentido horário: (a) sentido para o interior; (b) sentido para o exterior.

Fonte: Elaborada pelo autor.

coordenadas, os valores de X são selecionados ao longo do eixo X da grade; os valores de Y e Z , no entanto, têm seus valores selecionados para cada camada $x = k$ da grade seguindo as espirais retangulares, alternadamente do *voxel* central para as bordas e ao contrário. Deste modo, X , Y e Z mantêm sua associação, os valores de X ficam ordenados de modo majoritariamente não decrescente e valores de Y e de Z ficam ordenados com menor variação do que o método da seção 4.4. Uma vez a grade cúbica de projeção preenchida, o cálculo de momentos e reconstrução da superfície segue o procedimento original.

5 RECONSTRUÇÃO DE SUPERFÍCIES MOLECULARES EM ZERNIKE 3D

Neste capítulo é apresentada a avaliação de representações de superfícies moleculares em polinômios tridimensionais de Zernike. Tendo em vista a experiência prévia com implementações que calculavam resultados insatisfatórios, inicia-se o estudo de reconstrução de funções em Zernike 3D com a avaliação de ortonormalidade de duas definições de polinômios. Em seguida, são avaliadas a reconstrução usual de funções binárias para objetos volumétricos e a reconstrução de funções gerais. Finalmente, as duas alternativas para a representação de superfícies moleculares definidas por funções gerais (não binárias) no capítulo 4 são discutidas.

5.1 Avaliação de integração

Integração de Monte Carlo do método foi aplicada na equação 1.16 para verificar a ortonormalidade de polinômios de Zernike. O número de pontos de integração determinado pelo algoritmo *ball grid*¹¹⁵ foi de 2184897, e o índice n variou de 0 a 3. A tabela 6 mostra todos os resultados que diferiram do valor esperado de $\delta_{nn'}\delta_{ll'}\delta_{mm'}$ no intervalo definido para n . Valores na coluna *Canterakis* seguem a definição de [Canterakis](#)⁶⁰ (equação 4.1), e valores na coluna *Mak* seguem [Mak, Grandison e Morris](#)⁶¹ (equação 4.3). As menores diferenças entre os valores obtidos e esperados favorecem a segunda definição.

Confiança nestes testes poderia ser aumentada ainda mais. Os polinômios 3D de Zernike definidos por [Mak, Grandison e Morris](#)⁶¹ foram ainda submetidos a teste de ortonormalidade por integração adaptativa¹¹⁷ através do cálculo da equação 1.17, referente à ortonormalidade, pela biblioteca *cubature*.¹¹⁹ Como cálculos iniciais exauriram toda a memória disponível, o domínio de integração inicial foi dividido em porções de igual tamanho e cada um deles manipulado separadamente pelo integrador. Após soma dos resultados de cada porção, valores corretos foram obtidos para $\delta_{nn'}\delta_{ll'}\delta_{mm'}$ com erro absoluto inferior a 10^{-12} . Estes resultados, combinados com os testes das funções de harmônicos esféricos da seção 3.2 firmaram confiança satisfatória na implementação dos polinômios 3D de Zernike.

5.2 Reconstrução de funções binárias

No caso de formas tridimensionais, uma função binária representando um objeto voxelizado é usada na projeção em Zernike 3D, cujos valores reconstruídos são comparados a um limiar para se remover ruído.⁶¹ A implementação de referência de [Novotni e Klein](#),¹¹⁴ que segue esta abordagem, falhou em reconstruir qualquer uma das funções testadas, até mesmo uma função contínua $f(\mathbf{x}) = 1$. Entretanto, o programa de *docking* de proteínas LZerD¹²²

Tabela 6 – Resultados de integração para duas definições de polinômios tridimensionais de Zernike.

n	l	m	n'	l'	m'	$\delta_{nn'}\delta_{ll'}\delta_{mm'}$	Canterakis	Mak
0	0	0	0	0	0	1	0.333333333	1.000000000
0	0	0	2	0	0	0	-1.527472008	-0.000159671
1	1	-1	1	1	-1	1	0.333310105	0.999930314
1	1	0	1	1	0	1	0.333310105	0.999930314
1	1	1	1	1	1	1	0.333310105	0.999930314
2	0	0	0	0	0	0	-1.527472008	-0.000159671
2	0	0	2	0	0	1	7.332798779	0.999859747
2	2	-2	2	2	-2	1	0.333273428	0.999820283
2	2	-1	2	2	-1	1	0.333273428	0.999820283
2	2	0	2	2	0	1	0.333306931	0.999920793
2	2	1	2	2	1	1	0.333273428	0.999820283
2	2	2	2	2	2	1	0.333306931	0.999920793
3	1	-1	1	1	-1	0	-2.235834085	-0.000234207
3	1	-1	3	1	-1	1	15.331170266	0.999788898
3	1	-1	3	3	-3	0	0.000124756	-0.000062153
3	1	-1	3	3	-1	0	0.000096635	-0.000048143
3	1	0	1	1	0	0	-2.235834085	-0.000234207
3	1	0	3	1	0	1	15.331170266	0.999788898
3	1	0	3	3	0	0	-0.000157805	0.000078618
3	1	1	1	1	1	0	-2.235834085	-0.000234207
3	1	1	3	1	1	1	15.331170266	0.999788898
3	1	1	3	3	1	0	0.000096635	-0.000048143
3	1	1	3	3	3	0	-0.000124756	0.000062153
3	3	-3	1	1	-1	0	-0.000015509	-0.000046527
3	3	-3	3	1	-1	0	0.000124756	-0.000062153
3	3	-3	3	3	-3	1	0.333267805	0.999803416
3	3	-3	3	3	-1	0	0.000116698	0.000350093
3	3	-2	3	3	-2	1	0.333418492	1.000255476
3	3	-1	1	1	-1	0	-0.000012013	-0.000036040
3	3	-1	3	1	-1	0	0.000096635	-0.000048143
3	3	-1	3	3	-3	0	0.000116698	0.000350093
3	3	-1	3	3	-1	1	0.333207543	0.999622629
3	3	0	1	1	0	0	0.000019618	0.000058853
3	3	0	3	1	0	0	-0.000157805	0.000078618
3	3	0	3	3	0	1	0.333358199	1.000074597
3	3	1	1	1	1	0	-0.000012013	-0.000036040
3	3	1	3	1	1	0	0.000096635	-0.000048143
3	3	1	3	3	1	1	0.333207543	0.999622629
3	3	1	3	3	3	0	-0.000116698	-0.000350093
3	3	2	3	3	2	1	0.333117149	0.999351448
3	3	3	1	1	1	0	0.000015509	0.000046527
3	3	3	3	1	1	0	-0.000124756	0.000062153
3	3	3	3	3	1	0	-0.000116698	-0.000350093
3	3	3	3	3	3	1	0.333267805	0.999803416

Fonte: Elaborada pelo autor.

Figura 52 – Superfície molecular voxelizada (a) de 2K5X_A com resolução = 64^3 e reconstruções com (b) $N_{max} = 10$; (c) $N_{max} = 20$; (d) $N_{max} = 30$.

Fonte: Elaborada pelo autor.

usa esta abordagem de função binária para representação de superfícies moleculares, favorecendo a validade do algoritmo. A busca por implementações alternativas levou ao programa SASTBX.¹²⁰

A superfície molecular de 2K5X_A, calculada pelo programa EDTSurf conforme seção 4.2, foi voxelizada com resolução de 64, equivalente a uma grade retangular de $64^3 = 262144$ *voxels*, e é mostrada na figura 52 (a). A superfície foi projetada em Zernike 3D e reconstruída pelo programa SASTBX; as reconstruções são mostradas na figura 53 (b) para o parâmetro $N_{max} = 10$ (286 momentos), na figura 53 (c) para $N_{max} = 20$ (1771 momentos), e na na figura 53 (d) para $N_{max} = 30$ (5456 momentos). O caso com N_{max} igual a 30 exibe maior semelhança com a superfície original, ao passo que ajustar $N_{max} \geq 40$ produziu reconstruções completamente errôneas.

A função binária original da superfície de 2K5X_A e suas reconstruções são mostradas nas figuras 53 (a-d), em que a linha horizontal demarca o limiar, estabelecido como 1/3 do máximo valor reconstruído. A lista linear de *voxels* foi obtida percorrendo-se a grade retangular de projeção primeiramente na coordenada z , depois na coordenada y , e finalmente na coordenada x . Para $N_{max} \in \{10, 20, 30\}$, os valores reconstruídos não atingiram 50% do valor original dos *voxels* preenchidos, e com $N_{max} = 40$ a reconstrução foi completamente errônea. Novotni e Klein⁶³ comentam a necessidade de uso de ordens N_{max} de reconstrução até 20 para o provimento de bons descritores. Embora o uso de um limiar tenha resultado em reconstruções que capturam o aspecto geral do objeto de entrada, esta estratégia pode remover porções válidas de objetos reconstruídos e ao mesmo tempo introduzir informação alheia ao objeto, além de utilizar diferentes valores de limiar para diferentes ordens de reconstrução. Ademais, diferentes ordens de reconstrução usualmente requerem diferentes valores de limiar. O limiar igual a 1/3 do valor máximo foi estabelecido de forma empírica para este exemplo; a literatura não menciona uma estratégia para sua adoção. Grandison, Roberts e Morris¹²³ ilustraram a obtenção de diferentes superfícies para limiares distintos.

Figura 53 – Valores reconstruídos de 2K5X_A em voxelização binária com resolução $= 64^3$ por SASTBX: (a) $N_{max} = 10$; (b) $N_{max} = 20$; (c) $N_{max} = 30$; (d) $N_{max} = 40$.

Fonte: Elaborada pelo autor.

5.3 Reconstrução de funções gerais

Momentos de Zernike 3D foram usados para comparação de formas^{124,125} e para *docking* de proteínas.^{62,122} No entanto, apenas funções binárias foram consideradas. A estratégia de escolha de limiar sequer é mencionada. Sael et al.¹²⁴ utilizaram dois conjuntos de momentos de Zernike 3D para representar uma função não binária, um para os valores positivos e outro para os valores negativos. Já Zhu, Xiong e Kihara¹²⁶ representaram moléculas em Zernike 3D para buscar sítios de ligação a partir de um banco de dados de sítios conhecidos. Contudo, em ambos os trabalhos os momentos não foram usados para reconstrução da função original, mas sim para cálculo de descritores de Zernike a partir dos momentos Ω_{nl}^m para comparação de objetos 3D independentemente de alinhamento.¹²⁵

Nisius e Gohlke¹²¹ alegam ser capazes de representar uma função geral, não binária, em Zernike 3D. Seu programa foi avaliado para se tentar reconstruir funções gerais não binárias. As figuras 54 (a-c) mostram reconstruções com $N_{max} = 10, 15$ e 20 de uma função exemplo, fornecida pelo programa como um conjunto de valores discretos, bem como suas reconstruções. Ainda que a correlação tenha sido elevada, superior a 70%, há considerável divergência nas reconstruções de valores originais maiores do que 0,5. Reconstruções com ordem igual a 30 resultaram em valores sem correspondência com os originais.

O código de SASTBX foi retrabalhado, em termos de entrada e saída de valores sem alteração nas rotinas de cálculo, para lidar com funções não binárias gerais e foi testado com as funções de coordenada X , Y e Z novamente de 2K5X_A em grid de resolução 40. A resolução menor deste teste foi escolhida para permitir visualização mais clara dos resultados; reconstruções com resolução 64 apresentaram comportamento semelhante. As figuras 55, 56 e 57 apresentam os resultados para X , Y e Z , respectivamente, com $N_{max} = 10$ (a), 20 (b) e 30 (c). O formato geral das funções de entrada pode ser visto, mas também estão evidentes vários *voxels* posicionados incorretamente entre os valores das coordenadas no topo e *voxels* vazios na base.

Uma vez que *voxels* vazios não adicionam informações ao formato do objeto (neste caso em que o formato é separado em funções coordenadas e não representado de forma binária), eles foram omitidos dos cálculos. No entanto, o código retrabalhado de SASTBX não o permitia. O programa de Nisius e Gohlke¹²¹ foi então usado para reconstruir com $N_{max} = 20$ as funções X , Y e Z de 2K5X_A sem voxelização prévia, ou seja, as funções foram diretamente obtidas das coordenadas de superfície calculadas por EDTSurf e foram voxelizadas internamente pelo programa de reconstrução em Zernike 3D. Os resultados são mostrados na figura 58. A função reconstruída X (figura 58 (a)) precisou ser multiplicada por 100 se aproximar dos valores originais (figura 58 (b)). As funções Y e Z , por sua vez, não apresentaram similaridades óbvias com suas partes originais (figuras 58 (c) e 58 (d)). Embora as funções Y e Z e a função de exemplo sejam altamente oscilatórias, Y e Z não é possível identificar facilmente um formato geral.

Figura 54 – Reconstruções de uma função de exemplo pelo programa de [Nisius e Gohlke](#)¹²¹ usando (a) $N_{max} = 10$, (b) $N_{max} = 15$, e (c) $N_{max} = 20$.

Fonte: Elaborada pelo autor.

Figura 55 – Valores de *voxel* para as coordenadas X da superfície de 2K5X_A calculados pelo programa SASTBX.

Fonte: Elaborada pelo autor.

Figura 56 – Valores de *voxel* para as coordenadas Y da superfície de 2K5X_A calculados pelo programa SASTBX.

Fonte: Elaborada pelo autor.

Figura 57 – Valores de *voxel* para as coordenadas Z da superfície de 2K5X_A calculados pelo programa SASTBX.

Fonte: Elaborada pelo autor.

Figura 58 – Reconstruções dos componentes coordenados de 2K5X_A pelo programa de Nisius e Gohlke¹²¹ com $N_{max} = 20$: (a) valores de X; (b) valores de X multiplicados por 1000 após reconstrução; (c) valores de Y; (d) valores de Z.

Fonte: Elaborada pelo autor.

5.4 Representação em Zernike 3D com reordenação axial

Os resultados anteriores esclareceram as principais dificuldades relativas a reconstruções em Zernike 3D: a baixa cobertura de pontos na bola unitária e a baixa suavidade da função projetada, já que os algoritmos tentaram aproximar todos os valores presentes na bola unitária, até mesmo os referentes a *voxels* vazios, e calcularam valores intermediários quando houve acentuada variação entre um valor de entrada e seus vizinhos. Os momentos geométricos, que são usados no cálculo dos momentos de Zernike, já foram classificados como severamente instáveis em reconstruções com ordens elevadas.⁶³ Para Grandison, Roberts e Morris,¹²³ a representação em Zernike 3D é capaz de representar funções tridimensionais gerais, desde que variem “razoavelmente suavemente” no espaço 3D. Ademais, suas reconstruções com ordem máxima 20 e grade de 64^3 *voxels* assumidamente omitem detalhes finos dos objetos reconstruídos.¹²³ Sit et al.¹²⁷ classificaram como uma desvantagem a ortogonalidade de polinômios de Zernike 3D ser válida apenas na

Figura 59 – (a) Superfície não binária voxelizada de 2K5X_A com resolução $= 100^3$. (b) reconstrução de 2K5X_A na abordagem de reordenação axial com $N_{max} = 15$; (c) malha triangular calculada por Powercrust a partir de (b).

Fonte: Elaborada pelo autor.

bola unitária e propuseram uma reortogonalização para torná-la válida em um domínio retangular; no entanto, isto por si só não elimina os *voxels* vazios que são alheios à função de interesse.

Estas considerações inspiraram a consideração dos valores de entrada não como uma grade cúbica que contém a esfera unitária, mas sim como uma mera lista unidimensional de valores, de modo similar ao qual os valores são introduzidos no programa de Nisius e Gohlke,¹²¹ culminando na proposição do método da seção 4.5. Para avaliá-lo, a superfície da figura 59 (a) foi, conforme tal procedimento, deformada na bola unitária e teve seus valores para Y e Z reordenados para torná-los monotonicamente não decrescentes, ou seja, $\forall a, b, a \leq b \implies f(a) \leq f(b), f \in \{Y, Z\}$. A resolução igual a 100 do objeto original foi escolhido para permitir mais detalhes do que os usuais 64^3 dos resultados de Novotni e Klein.¹¹⁴

As figuras 60, 61 e 62 mostram funções de coordenada X , Y e Z originais e seus valores reconstruídos para $N_{max} = 20$, com um total de 1771 momentos Ω_{nl}^m . No caso de Y e Z , os valores fornecidos como entrada ao método de reconstrução são mostrados em (b), e os valores da função na ordem original da superfície são mostrados em (a). A acentuada queda na ponta direita das curvas são causadas por *voxels* não preenchidos da grade de reconstrução devido a insuficiência de pontos da função original, o que pode ser corrigido pela replicação do último valor da função em tais *voxels*. O método facilmente reproduz as funções de entrada e oferece mais detalhes na superfície reconstruída para os mesmos valores de N_{max} , vistos na figura 59 (b), do que outras abordagem de Zernike, sem precisar lançar mão de limiar, como fizeram Novotni e Klein,¹¹⁴ ou usar diferentes conjuntos de momentos para valores positivos e negativos da função, como fizeram Sael et al.¹²⁴ Em um computador com números de ponto flutuante de 64 bits (8 bytes) e inteiros de 32 bits (4 bytes), a superfície de exemplo 2K5X_A representada desta forma requeria

Figura 60 – Valores de *voxel* da coordenada X da superfície de 2K5X_A na abordagem de mapeamento na bola com reordenação axial.

Fonte: Elaborada pelo autor.

$1771 \times 8 = 14168$ bytes = 18,836 KB de armazenamento para os momentos e 291 KB para o mapa; isto é uma ordem de magnitude menos espaço de memória do que os 2801 KB requeridos pela malha de entrada original. O pior caso para uso de memória deste método se dá quando o objeto de entrada ocupa todos os *voxels* na bola unitária, ou seja, o objeto é a própria bola unitária. Seja G a resolução da grade de reconstrução que contém a bola unitária, formando uma grade cúbica de lado igual a 2 e G^3 *voxels*. Como o volume da bola unitária é $4\pi/3$ e o volume da grade cúbica é $2^3 = 8$, há $\frac{4\pi G^3}{3} \frac{1}{8} \approx 0,5236 \times G^3$ *voxels* na bola unitária, ou cerca de $0,5 \times G^3$ *voxels*, que ocupam cerca de $8 \times 0,5 \times G^3 = 4 \times G^3$ bytes de memória. Novotni e Klein¹¹⁴ tiveram seus melhores resultados de suas restaurações de funções volumétricas binárias em uma grade com resolução de 64^3 . A superfície de 2K5X_A calculada por EDTSurf possui 59748 vértices, resultando numa grade de reconstrução com $40^3 = 64000$ *voxels*. Entretanto, grades mais finas capturam mais detalhes e são necessários para modelar corretamente a superfícies de moléculas maiores, que possuem superfície molecular de maior área e, portanto, com mais pontos a serem posicionados na grade de reconstrução.

Ainda que o aspecto geral da superfície de entrada tenha sido capturado, o mapa deste método precisou ser mantido e manipulado para adequada restauração das funções, ao contrário do mapa esférico da representação em harmônicos esféricos dos capítulos 2 e 3. Desta forma, uma porção de dados alheia à representação em Zernike 3D precisa ser armazenada e manipulada para se recuperar o objeto representado. É válido apenas

(a)

(b)

Figura 61 – Valores de *voxel* da coordenada Y da superfície de 2K5X_A na abordagem de mapeamento na bola: (a) usando a ordem original da superfície molecular; (b) usando a ordem do mapeamento

Fonte: Elaborada pelo autor.

(a)

(b)

Figura 62 – Valores de *voxel* da coordenada Z da superfície de 2K5X_A na abordagem de mapeamento na bola: (a) usando a ordem original da superfície molecular; (b) usando a ordem do mapeamento.

Fonte: Elaborada pelo autor.

Figura 63 – (a) Função X da superfície de 2K5X_A visitada seguindo uma espiral. (b) Comparação entre função original X e sua reconstrução com $N_{max} = 30$.

Fonte: Elaborada pelo autor.

para as funções coordenadas que descrevem a superfície, mas não para propriedades gerais definidas como uma única função na superfície. Por exemplo, seria despropósito tentar reordenar os valores de energia potencial eletrostática seguindo este esquema. Fórmulas parecidas com as de [Duncan e Olson⁴⁶](#) para cálculos de curvaturas da superfície não farão sentido neste método, posto que a função representada pelos momentos de Zernike não é a mesma função da superfície, ainda que o conjunto de seus valores seja o mesmo.

5.5 Representação em Zernike 3D com reordenação espiral

As figuras 63 (a), 64 (a) e 65 (a) mostram as funções coordenadas X , Y e Z de 2K5X_A calculadas por EDTSurf, voxelizadas em grade de resolução 64 e listadas segundo a espiral da seção 4.6. O comportamento de X é o mesmo das análises dos parágrafos anteriores, ao passo que Y e Z apresentam acentuada oscilação. As respectivas reconstruções em Zernike 3D com $N_{max} = 30$ são mostradas nas figuras 63 (b), 64 (b) e 65 (b). Conforme esperado devido a seu comportamento geral crescente, ainda que com curtas regiões de baixa oscilação que se assemelham a degraus, a reconstrução de X teve acentuada correspondência com os valores originais. Ainda que a correlação das reconstruções de Y e Z tenha sido elevada, de cerca de 88%, muitos valores divergiram do esperado, em um comportamento análogo à da função de exemplo no programa de [Nisius e Gohlke¹²¹](#) da seção 5.3. Após composição das três funções em coordenadas cartesianas, a reconstrução da figura 66 (a) foi insatisfatória, pois os pontos reconstruídos não corresponderam ao objeto volumétrico original de 2K5X_A da figura 66 (b).

Figura 64 – (a) Função Y da superfície de 2K5X_A visitada seguindo uma espiral. (b) Comparação entre função original Y e sua reconstrução com $N_{max} = 30$.

Fonte: Elaborada pelo autor.

Figura 65 – (a) Função Z da superfície de 2K5X_A visitada seguindo uma espiral. (b) Comparação entre função original Z e sua reconstrução com $N_{max} = 30$.

Fonte: Elaborada pelo autor.

Figura 66 – (a) Superfície molecular voxelizada de 2K5X_A; (b) sua reconstrução com $N_{max} = 30$ no método de ordenação espiral.

Fonte: Elaborada pelo autor.

A tentativa de evitar oscilações bruscas da função através de listagem de pontos seguindo uma espiral não foi bem sucedida. Conforme a equação 4.13, ordens de reconstrução N_{max} superiores são necessárias para se aumentar a fidelidade das reconstruções. Grandison, Roberts e Morris,¹²³ por exemplo, reconhecem a ausência de detalhes mais finos em suas reconstruções no algoritmo tradicional de Zernike 3D com ordem máxima 20 e grade de 64^3 *voxels*, ao passo que Liu et al.¹²⁸ indicam a necessidade de uso de ordem de reconstrução $N_{max} \geq 30$ para “macromoléculas típicas”. No entanto, a quantidade de momentos é $\mathcal{O}((N_{max})^3)$, ou seja, cresce limitada a $(N_{max})^3$ com o aumento da ordem de reconstrução. O uso de dados da reconstrução insatisfatória da figura 66 é de 5456 momentos para cada função, totalizando 16368 momentos, equivalentes a 130944 bytes ou aproximadamente 128KB de dados num computador com números de ponto flutuante de 8 bytes. Em comparação, a reconstrução de qualidade superior da superfície da mesma molécula 2K5X_A com 21 bandas de harmônicos esféricos da figura 26 usa apenas $3 \times 441 = 1323$ coeficientes, somando 11KB de dados. Como a quantidade de coeficientes de harmônicos esféricos é $\mathcal{O}((L_{max})^2)$, a representação com listagem espiral é desvantajosa em relação à de harmônicos esféricos quanto à quantidade de dados que precisam ser manipulados. Ademais, em algumas reconstruções em Zernike 3D com ordem igual ou superior a 40 e implementação de SASTBX, como o exemplo da figura 53 (d), e com ordem igual a 30 com a implementação de Nisius e Gohlke,¹²¹ manifestaram explosão de valores e indicando necessidade de revisão cuidadosa de seus algoritmos. De todo modo, este método se aplica apenas em funções coordenadas que descrevam a geometria de um objeto volumétrico, não sendo adequado portanto para funções gerais que definam propriedades na superfície molecular.

A elevada quantidade de momentos necessária para representar funções gerais em Zernike 3D nos quatro métodos estudados pode tornar esta abordagem menos conveniente do que a de harmônicos esféricos quando reconstruções altamente detalhadas forem requeridas, reduzindo, portanto, a vantagem de capacidade natural de representação de

túneis em Zernike 3D sem demandar mapeamento esférico no contexto de atracamento de proteínas.

6 CONCLUSÃO

A representação de superfícies moleculares em elevado nível de detalhes é um requisito importante para o *docking* de proteínas. O tratamento implícito de flexibilidade conformacional através de um *ensemble* de conformações requer uma representação compacta de superfícies moleculares para reduzir a demanda por recursos computacionais. Representações baseadas em harmônicos esféricos têm sido usadas para este propósito por permitirem descrever superfícies com pequeno volume de dados e controle de nível de detalhamento e por possuírem rotação eficiente. A representação paramétrica em harmônicos esféricos e a representação em polinômios tridimensionais de Zernike descrevem, a seu modo, superfícies com relevo acidentado. No entanto, enquanto a primeira é limitada a superfícies sem túneis, devido à etapa intermediária de mapeamento esférico, a segunda é utilizada para reconstruções de funções binárias. Em ambos os casos a qualidade da representação é passível de melhoria, com consequentes benefícios para o *docking* de proteínas.

Este estudo explorou os mecanismos subjacentes a ambas as representações para propor uma combinação de algoritmos estável e abrangente. Na representação paramétrica em harmônicos esféricos, a distribuição de amostras da função representada pela superfície esférica foi identificada como o principal fator determinante de qualidade de reconstruções, com maior regularidade favorecendo reconstruções menos distorcidas. A combinação entre os estágios de cálculo de malha triangular de superfície molecular e de mapeamento esférico influenciou diretamente a distribuição de pontos, de modo que nenhum dos componentes separadamente é capaz de garantir regularidade adequada de pontos pelo domínio de representação. O fechamento de túneis generalizou a abordagem para superfícies de topologia arbitrária sem introduzir acentuadas distorções. Ademais, as distorções resultantes foram locais, afetando apenas a vizinhança dos túneis. Uma análise de similaridade aplicada a *ensemble* de conformações capturou os detalhes finos e o aspecto geral de acordo com a medida de variabilidade empregada. A combinação de refinamento sucessivo de qualidade com decomposição espectral reduziu grandemente a quantidade de dados a manipular após obtenção de coeficientes de harmônicos esféricos e permitiu reconstruir superfícies com grau ajustável de detalhamento.

Na representação em polinômios de Zernike 3D foi detectada alta sensibilidade a mudança de valores nas funções projetadas na qualidade da reconstrução, prejudicando a descrição de superfícies em razão de valores vazios no domínio de representação e impedindo descrição adequada de funções gerais definidas na superfície de objetos volumétricos pela abordagem tradicional. Uma proposta alternativa reduziu a presença de valores alheios à função de interesse e permitiu reproduzir superfície molecular. No entanto, a elevada

quantidade de coeficientes necessários para reconstruções de alta qualidade torna esta abordagem menos atrativa do que a representação paramétrica em harmônicos esféricos.

A compilação resultante de métodos tem aplicação indicada em metodologias de atracamento de proteínas. Neste contexto, a combinação de ferramentas permite representação compacta em alta qualidade de superfícies moleculares que auxilia detecção, em baixo nível de detalhes, de orientações mais favoráveis das moléculas e eliminação de orientações desfavoráveis nas etapas seguintes de busca com superfícies mais detalhadas. Há interesse da representação em outras áreas além de *docking* de proteínas, como, por exemplo, engenharia civil¹¹² e análise de imagens médicas.¹¹³ Melhorias na representação paramétrica podem beneficiar outras áreas de conhecimento.

Investigações futuras são propostas para se estabelecer seleção automática de quantidade de coeficientes comuns a usar para um *ensemble* de conformações e refinamento de identificação de similaridade entre superfícies. Operações de rotação e de detecção de colisões de superfícies descritas de forma paramétrica em harmônicos esféricos podem compor o estágio de dinâmica de corpos rígidos de uma aplicação de atracamento de proteínas lançando mão de rotação mais eficiente de harmônicos esféricos do que de pontos no espaço cartesiano.^{129,130}

REFERÊNCIAS

- 1 RUSSELL, R. B. et al. A structural perspective on protein–protein interactions. **Current Opinion in Structural Biology**, v. 14, n. 3, p. 313–324, 2004.
- 2 BRANDEN, J. T. C. **Introduction to protein structure**. 2nd. ed. New York: Garland Publishing, Inc., 1998.
- 3 CREIGHTON, T. E. **Proteins: structures and molecular properties**. 2nd. ed. New York: W. H. Freeman and Company, 1993.
- 4 PETSKO, G. A.; RINGE, D. **Protein structure and function**. London: New Science Press Ltd, 2004. 195 p.
- 5 ALBERTS, B. The cell as a collection of protein machines: preparing the next generation of molecular biologists. **Cell**, v. 92, n. 3, p. 291–294, 1998.
- 6 WÜTHRICH, K. The way to NMR structures of proteins. **Nature Structural Biology**, v. 8, n. 11, p. 923–925, 2001.
- 7 CAVALLI, A. et al. Protein structure determination from NMR chemical shifts. **Proceedings of the National Academy of Sciences**, v. 104, n. 23, p. 9615–9620, 2007.
- 8 BONOMI, M.; PELLARIN, R.; VENDRUSCOLO, M. Simultaneous determination of protein structure and dynamics using cryo-electron microscopy. **Biophysical Journal**, v. 114, n. 7, p. 1604–1613, 2018.
- 9 BLUNDELL, T. L.; JOHNSON, L. N. **Protein crystallography**. New York: Academic Press, 1976. 565 p.
- 10 WODAK, S. J.; JANIN, J. Computer analysis of protein-protein interaction. **Journal of Molecular Biology**, v. 124, n. 2, p. 323–342, 1978.
- 11 VAKSER, I. A.; KUNDROTAS, P. Predicting 3d structures of protein-protein complexes. **Current Pharmaceutical Biotechnology**, v. 9, n. 2, p. 57–66, 2008.
- 12 MCCAMMON, J. A. Theory of biomolecular recognition. **Current Opinion in Structural Biology**, v. 8, n. 2, p. 245–249, 1998.
- 13 DILL, K. A.; MACCALLUM, J. L. The protein-folding problem, 50 years on. **Science**, v. 338, n. 6110, p. 1042–1046, 2012.
- 14 KUHL, F. S.; CRIPPEN, G. M.; FRIESEN, D. K. A combinatorial algorithm for calculating ligand binding. **Journal of Computational Chemistry**, v. 5, n. 1, p. 24–34, 1984.
- 15 SHAW, D. E. et al. Atomic-level characterization of the structural dynamics of proteins. **Science**, v. 330, n. 6002, p. 341–346, 2010.
- 16 LANE, T. J. et al. To milliseconds and beyond: challenges in the simulation of protein folding. **Current Opinion in Structural Biology**, v. 23, n. 1, p. 58–65, 2013.

- 17 CUZZOLIN, A. et al. Deciphering the complexity of ligand–protein recognition pathways using supervised molecular dynamics (sumd) simulations. **Journal of Chemical Information and Modeling**, v. 56, n. 4, p. 687–705, 2016.
- 18 GIOIA, D. et al. Dynamic docking: a paradigm shift in computational drug discovery. **Molecules**, v. 22, n. 11, 2017.
- 19 LEVINTHAL, C. et al. Hemoglobin interaction in sickle cell fibers. I: theoretical approaches to the molecular contacts. **Proceedings of the National Academy of Sciences**, v. 72, n. 4, p. 1330–1334, 1975.
- 20 RITCHIE, D. W. Recent progress and future directions in protein-protein docking. **Current Protein and Peptide Science**, v. 9, n. 1, p. 1–15, 2008.
- 21 WEI, G. et al. Protein ensembles: how does nature harness thermodynamic fluctuations for life? the diverse functional roles of conformational ensembles in the cell. **Chemical Reviews**, v. 116, n. 11, p. 6516–6551, 2016.
- 22 FRAUENFELDER, H.; SLIGAR, S.; WOLYNES, P. The energy landscapes and motions of proteins. **Science**, v. 254, n. 5038, p. 1598–1603, 1991.
- 23 TOKURIKI, N.; TAWFIK, D. S. Protein dynamism and evolvability. **Science**, v. 324, n. 5924, p. 203–207, 2009.
- 24 MONTALVAO, R. W.; SIMONE, A. D.; VENDRUSCOLO, M. Determination of structural fluctuations of proteins from structure-based calculations of residual dipolar couplings. **Journal of Biomolecular NMR**, v. 53, n. 4, p. 281–292, 2012.
- 25 B-RAO, C.; SUBRAMANIAN, J.; SHARMA, S. D. Managing protein flexibility in docking and its applications. **Drug Discovery Today**, v. 14, n. 7, p. 394–400, 2009.
- 26 LENSINK, M. F.; VELANKAR, S.; WODAK, S. J. Modeling protein-protein and protein-peptide complexes: CAPRI 6th edition. **Proteins**, v. 85, n. 3, p. 359–377, 2016.
- 27 BAHAR, A.-D.; MARC, P.; NURIT, H. Accurate prediction of docked protein structure similarity. **Journal of Computational Biology**, v. 22, n. 9, p. 892–904, 2015.
- 28 SUKHWAL, A.; SOWDHAMINI, R. PPCheck: a webserver for the quantitative analysis of protein-protein interfaces and prediction of residue hotspots. **Bioinformatics and Biology Insights**, v. 9, p. 141–151, 2015.
- 29 CHERFILS, J.; JANIN, J. Protein docking algorithms: simulating molecular recognition. **Current Opinion in Structural Biology**, v. 3, n. 2, p. 265–269, 1993.
- 30 MOREIRA, I. S.; FERNANDES, P. A.; RAMOS, M. J. Protein–protein docking dealing with the unknown. **Journal of Computational Chemistry**, v. 31, n. 2, p. 317–342, 2010.
- 31 BONVIN, A. M. Flexible protein–protein docking. **Current Opinion in Structural Biology**, v. 16, n. 2, p. 194–200, 2006.
- 32 MONTALVAO, R. W. et al. Structure determination of protein-protein complexes using NMR chemical shifts: case of an endonuclease colicin-immunity protein complex. **Journal of the American Chemical Society**, v. 130, n. 47, p. 15990–15996, 2008.

-
- 33 BERMAN, H. M. et al. The Protein Data Bank. **Nucleic Acids Research**, v. 28, n. 1, p. 235–242, 2000.
- 34 PROTEIN Data Bank contents guide: atomic coordinate entry format description. World Wide Protein Data Bank, 2012. Disponível em: <<http://www ww p d b . o r g / d o c u m e n t a t i o n / f i l e - f o r m a t - c o n t e n t / f o r m a t 3 3 / v 3 . 3 . h t m l>>. Acesso em: 17 abr. 2018.
- 35 LEE, B.; RICHARDS, F. M. The interpretation of protein structures: estimation of static accessibility. **Journal of Molecular Biology**, v. 55, n. 3, p. 379–400, 1971.
- 36 RITCHIE, D. W. **Parametric protein shape recognition**. 1998. Tese (Doutorado) — University of Aberdeen, Aberdeen, 1998.
- 37 NICOLAU, D. V. et al. Protein molecular surface mapped at different geometrical resolutions. **PLoS ONE**, v. 8, n. 3, p. 1–15, 2013.
- 38 KATCHALSKI-KATZIR, E. et al. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. **Proceedings of the National Academy of Sciences**, v. 89, n. 6, p. 2195–2199, 1992.
- 39 CHUNG, M. K. et al. Encoding cortical surface by spherical harmonics. **Statistica Sinica**, v. 18, n. 2008, p. 1269–1291, 2008.
- 40 SHEN, L.; MAKEDON, F. Spherical mapping for processing of 3D closed surfaces. **Image and Vision Computing**, v. 24, n. 7, p. 743–761, 2006.
- 41 BRECHBÜHLER, C.; GERIG, G.; KÜBLER, O. Parametrization of closed surfaces for 3-D shape description. **Computer Vision and Image Understanding**, v. 61, n. 2, p. 154–170, 1995.
- 42 MOUSA, M.-H. et al. Toward an efficient triangle-based spherical harmonics representation of 3D objects. **Computer Aided Geometric Design**, v. 25, n. 8, p. 561–575, 2008.
- 43 SU, Z.; COPPENS, P. Rotation of real spherical harmonics. **Acta Crystallographica Section A**, v. 50, n. 5, p. 636–643, 1994.
- 44 PARIMAL, S.; CRAMER, S. M.; GARDE, S. Application of a spherical harmonics expansion approach for calculating ligand density distributions around proteins. **The Journal of Physical Chemistry B**, v. 118, n. 46, p. 13066–13076, 2014.
- 45 GREEN, R. **Spherical harmonic lighting**: the gritty details. 2003. Disponível em: <<http://silviojemma.com/public/papers/lighting/spherical-harmonic-lighting.pdf>>. Acesso em: 14 ago. 2018.
- 46 DUNCAN, B. S.; OLSON, A. J. Approximation and characterization of molecular surfaces. **Biopolymers**, v. 33, n. 2, p. 219–229, 1993.
- 47 LEICESTER, S.; FINNEY, J.; BYWATER, R. A quantitative representation of molecular surface shape. I: theory and development of the method. **Journal of Mathematical Chemistry**, v. 16, n. 1, p. 315–341, 1994.

- 48 RITCHIE, D. W.; KEMP, G. J. Protein docking using spherical polar fourier correlations. **Proteins**, v. 39, n. 2, p. 178–194, 2000.
- 49 KŘIVÁNEK, J. et al. Fast approximation to spherical harmonics rotation. In: SPRING CONFERENCE ON COMPUTER GRAPHICS, 22., 2006, Častá-Papiernička. **Proceedings...** New York: ACM, 2006. p. 1–10.
- 50 ARFKEN, G. B.; WEBER, H. J. **Mathematical methods for physicists**. 6th. ed. Burlington: Elsevier Academic Press, 2005.
- 51 PRESS, W. H. et al. **Numerical recipes: the art of scientific computing**. 3rd. ed. New York: Cambridge University Press, 2007.
- 52 DENNERY, P. **Mathematics for physicists**. New York: Harper and Row, 1967.
- 53 GOODMAN, S. E. **Beginning topology**. Providence: American Mathematical Society, 2009.
- 54 BRECHBÜHLER, C.; GERIG, G.; KUBLER, O. Surface parametrization and shape description. In: THE INTERNATIONAL SOCIETY FOR OPTICAL ENGINEERING, VISUALIZATION IN BIOMEDICAL COMPUTING, 1., 1992, Chapel Hill. **Proceedings...** Bellingham: SPIE, 1992. p. 80–89.
- 55 NORLEDGE, B. V. et al. The tissue factor/factor VIIa/factor Xa complex: a model built by docking and site-directed mutagenesis. **Proteins**, v. 53, n. 3, p. 640–648, 2003.
- 56 RITCHIE, D. W.; KEMP, G. J. L. Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces. **Journal of Computational Chemistry**, v. 20, n. 4, p. 383–395, 1999.
- 57 ZERNIKE, F. Beugungstheorie des schneidenver-fahrens und seiner verbesserten form, der phasenkontrastmethode. **Physica**, v. 1, n. 7, p. 689–704, 1934.
- 58 LIAO, S. X.; PAWLAK, M. On the accuracy of Zernike moments for image analysis. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 20, n. 12, p. 1358–1364, 1998.
- 59 THAPALIYA, K.; KWON, G.-R. Identification and extraction of brain tumor from mri using local statistics of Zernike moments. **International Journal of Imaging Systems and Technology**, v. 24, n. 4, p. 284–292, 2014.
- 60 CANTERAKIS, N. 3D Zernike moments and Zernike affine invariants for 3D image analysis and recognition. In: SCANDINAVIAN CONFERENCE ON IMAGE ANALYSIS, 11., 1999, Kangerlussuaq. **Proceedings...** New York: Springer, 1999. p. 85–93.
- 61 MAK, L.; GRANDISON, S.; MORRIS, R. J. An extension of spherical harmonics to region-based rotationally invariant descriptors for molecular shape description and comparison. **Journal of Molecular Graphics and Modelling**, v. 26, n. 7, p. 1035–1045, 2008.
- 62 ESQUIVEL-RODRIGUEZ, J. et al. Pairwise and multimeric protein–protein docking using the LZerD program suite. In: _____. **Protein structure prediction**. New York: Springer, 2014. p. 209–234.

- 63 NOVOTNI, M.; KLEIN, R. Shape retrieval using 3D Zernike descriptors. **Computer-Aided Design**, v. 36, n. 11, p. 1047–1062, 2004.
- 64 MOONEY, J. Strategies for supporting application portability. **Computer**, v. 23, n. 11, p. 59–70, 1990.
- 65 TIOBE Index. 2018. Disponível em: <<https://www.tiobe.com/tiobe-index/>>. Acesso em: 26 Ago. 2018.
- 66 USING the gnu compiler collection (gcc). Free Software Foundation, Inc., 2013. Disponível em: <<https://gcc.gnu.org/onlinedocs/gcc-4.8.4/gcc/>>. Acesso em: 26 ago. 2018.
- 67 MARTIN, K.; HOFFMAN, B. An open source approach to developing software in a small organization. **IEEE Software**, v. 24, n. 1, p. 46–53, 2007.
- 68 RAHI, S. J.; SHARP, K. Mapping complicated surfaces onto a sphere. **International Journal of Computational Geometry & Applications**, v. 17, n. 4, p. 305–329, 2007.
- 69 SURFMAP: software for mapping protein surfaces to a sphere. 2006. Disponível em: <<http://crystal.med.upenn.edu/software.html>>. Acesso em: 26 ago. 2018.
- 70 LORIOT, S.; F.CAZALS; J.BERNAUER. ESBTL: efficient PDB parser and data structure for the structural and geometric analysis of biological macromolecules. **Bioinformatics**, v. 26, n. 8, p. 1127–1128, 2010.
- 71 CGAL user and reference manual. CGAL Editorial Board, 2018. Disponível em: <<http://doc.cgal.org/4.11.1/Manual/packages.html>>. Acesso em: 17 abr. 2018.
- 72 TSAI, J. et al. The packing density in proteins: standard radii and volumes. **Journal of Molecular Biology**, v. 290, n. 1, p. 253–266, 1999.
- 73 XU, D.; ZHANG, Y. Generating triangulated macromolecular surfaces by Euclidean distance transform. **PLoS ONE**, v. 4, n. 12, p. 1–11, 2009.
- 74 LORENSEN, W. E.; CLINE, H. E. Marching cubes: a high resolution 3D surface construction algorithm. **SIGGRAPH Computer Graphics**, v. 21, n. 4, p. 163–169, 1987.
- 75 DEY, T. K.; FAN, F.; WANG, Y. An efficient computation of handle and tunnel loops via reeb graphs. **ACM Transactions on Graphics**, v. 32, n. 4, p. 1–10, 2013.
- 76 POLYGON mesh processing: 4.11.1. CGAL Editorial Board, 2018. Disponível em: <<http://doc.cgal.org/4.11.1/Manual/packages.html#PkgPolygonMeshProcessingSummary>>. Acesso em: 17 abr. 2018.
- 77 SHEN, L.; KIM, S.; SAYKIN, A. J. Fourier method for large-scale surface modeling and registration. **Computers & Graphics**, v. 33, n. 3, p. 299–311, 2009.
- 78 SPHARM-MAT, version 1.0.0: spherical harmonic modeling and analysis toolkit. Indianapolis: Indiana University, 2010. Disponível em: <<http://www.iu.edu/~spharm/>>. Acesso em: 14 Ago. 2018.

- 79 MATLAB version 7.9.0: the language of technical computing. Natick: The MathWorks Inc., 2009. Disponível em: <https://www.mathworks.com/products/matlab.html>. Acesso em: 14 ago. 2018.
- 80 KAZHDAN, M.; SOLOMON, J.; BEN-CHEN, M. Can mean-curvature flow be modified to be non-singular? **Computer Graphics Forum**, v. 31, n. 5, p. 1745–1754, 2012.
- 81 WANG, C.; LIU, Z.; LIU, L. As-rigid-as-possible spherical parametrization. **Graphical Models**, v. 76, n. 5, p. 457–467, 2014.
- 82 DAVIS, T. A. Algorithm 915, SuiteSparseQR: multifrontal multithreaded rank-revealing sparse QR factorization. **ACM Transactions on Mathematical Software**, v. 38, n. 1, p. 1–8, 2011.
- 83 WANG, C. et al. Bijective spherical parametrization with low distortion. **Computers & Graphics**, v. 58, n. 1, p. 161–171, 2016.
- 84 HU, X.; FU, X. M.; LIU, L. Advanced hierarchical spherical parameterizations. **IEEE Transactions on Visualization and Computer Graphics**, v. 24, n. 6, p. 1930–1941, 2018.
- 85 C1 quadratic interpolation of meshes. Geometric Tools, LLC, 2008. Disponível em: <http://www.geometrictools.com/Documentation/C1QuadraticInterpolation.pdf>. Acesso em: 17 abr. 2018.
- 86 WILD magic engine 5.13. Geometric Tools, LLC, 2014. Disponível em: <http://www.geometrictools.com/Downloads/Downloads.html>. Acesso em: 17 abr. 2018.
- 87 STABLE spherical harmonics for large l. 2003. Disponível em: <http://numerical.recipes/forum/showthread.php?t=200>. Acesso em: 30 nov. 2018.
- 88 CAFLISCH, R. E. Monte Carlo and quasi-Monte Carlo methods. **Acta Numerica**, v. 7, p. 1–49, 1998.
- 89 ARVO, J. Stratified sampling of spherical triangles. In: ANNUAL CONFERENCE ON COMPUTER GRAPHICS AND INTERACTIVE TECHNIQUES, 22., 1995, Los Angeles. **Proceedings...** New York: ACM, 1995. p. 437–438.
- 90 TODHUNTER, I. **Spherical trigonometry**: for the use of colleges and schools. 5th. ed. London: MacMillan and Co., 1886.
- 91 SHEN, L.; CHUNG, M. Large-scale modeling of parametric surfaces using spherical harmonics. In: INTERNATIONAL SYMPOSIUM ON 3D DATA PROCESSING, VISUALIZATION, AND TRANSMISSION, 3., 2006, Chapel Hill. **Proceedings...** Washington: IEEE, 2006. p. 294–301.
- 92 ANDERSON, E. et al. **LAPACK users' guide**. 3rd. ed. Philadelphia: Society for Industrial and Applied Mathematics, 1999.
- 93 GNU scientific library. 2018. Disponível em: <https://www.gnu.org/software/gsl/>. Acesso em: 27 ago. 2018.

- 94 SWINBANK, R.; PURSER, R. J. Fibonacci grids: a novel approach to global modelling. **Quarterly Journal of the Royal Meteorological Society**, v. 132, n. 619, p. 1769–1793, 2006.
- 95 AMENTA, N.; CHOI, S.; KOLLURI, R. K. The power crust, unions of balls, and the medial axis transform. **Computational Geometry**, v. 19, n. 2, p. 127–153, 2001.
- 96 CIGNONI, P. et al. Meshlab: an open-source mesh processing tool. In: EUROGRAPHICS ITALIAN CHAPTER CONFERENCE, 6., 2008, Salerno. **Proceedings...** Geneve: The Eurographics Association, 2008. p. 129–136.
- 97 GRAY, A. **Modern differential geometry of curves and surfaces**. Boca Raton: CRC Press, Inc., 1993.
- 98 ESTIMATION of local differential properties of point-sampled surfaces. CGAL Editorial Board, 2018. Disponível em: <http://doc.cgal.org/4.11.1/Jet_fitting_3/index.html#Chapter_Estimation_of_Local_Differential_Properties_of_Point-Sampled_Surfaces>. Acesso em: 17 abr. 2018.
- 99 BUTT, H.; GRAF, K.; KAPPL, M. **Physics and chemistry of interfaces**. Weinheim: Wiley-VCH, 2003.
- 100 HILDEBRANDT, A. et al. BALL - biochemical algorithms library 1.3. **BMC Bioinformatics**, v. 11, n. 1, p. 531, 2010.
- 101 LEYS, C. et al. Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. **Journal of Experimental Social Psychology**, v. 49, n. 4, p. 764–766, 2013.
- 102 DAUBECHIES, I. **Ten lectures on wavelets**. Philadelphia: Society for Industrial and Applied Mathematics, 1992.
- 103 VREVEN, T. et al. Updates to the integrated protein-protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. **Journal of Molecular Biology**, v. 427, n. 19, p. 3031–3041, 2015.
- 104 KRISHNAMURTHY, V.; LEVOY, M. Fitting smooth surfaces to dense polygon meshes. In: ANNUAL CONFERENCE ON COMPUTER GRAPHICS AND INTERACTIVE TECHNIQUES, 23., 1996, New Orleans. **Proceedings...** New York: ACM, 1996. p. 313–324.
- 105 MATHEMATICA version 9.0.1.0: sistema para computação técnica. Champaign: Wolfram Research, Inc., 2013. Disponível em: <<https://www.wolfram.com/mathematica/>>. Acesso em: 14 Ago. 2018.
- 106 GROSS, M.; STANIFORTH, A. Cubic-spline interpolation on a non-uniform latitude–longitude grid: achieving cross- and circum-polar continuity. **Atmospheric Science Letters**, v. 11, n. 3, p. 229–238, 2010.
- 107 BLENDER version 2.79b: free and open 3D creation software. Amsterdam: Blender Foundation, 2017. Disponível em: <<https://www.blender.org/>>. Acesso em: 14 Ago. 2018.

- 108 SCHMIDT, R.; SINGH, K. Meshmixer: an interface for rapid mesh composition. In: ACM SIGGRAPH TALKS, 37., 2010, Los Angeles. **Proceedings...** New York: ACM, 2010. p. 1–1.
- 109 BENSEDDIK, H.-E. et al. Direct method for rotation estimation from spherical images using 3D mesh surfaces with SPHARM representation. **Journal of Visual Communication and Image Representation**, v. 40, n. 1, p. 708–720, 2016.
- 110 MOUSA, M. et al. Efficient spherical harmonics representation of 3D objects. In: PACIFIC CONFERENCE ON COMPUTER GRAPHICS AND APPLICATIONS, 15., 2007, Maui. **Proceedings...** Washington: IEEE, 2007. p. 248–255.
- 111 SHIMAZAKI, H.; SHINOMOTO, S. A method for selecting the bin size of a time histogram. **Neural Computation**, v. 19, n. 6, p. 1503–1527, 2007.
- 112 SU, D.; YAN, W. 3D characterization of general-shape sand particles using microfocus X-ray computed tomography and spherical harmonic functions, and particle regeneration using multivariate random vector. **Powder Technology**, v. 323, n. 1, p. 8–23, 2018.
- 113 MIRAS, J. R. de et al. Complexity analysis of cortical surface detects changes in future Alzheimer’s disease converters. **Human Brain Mapping**, v. 38, n. 12, p. 5905–5918, 2017.
- 114 NOVOTNI, M.; KLEIN, R. 3D Zernike descriptors for content based shape retrieval. In: ACM SYMPOSIUM ON SOLID MODELING AND APPLICATIONS, 8., 2003, Seattle. **Proceedings...** New York: ACM, 2003. p. 216–225.
- 115 BALL_GRID - grid points within a 3D ball. 2017. Disponível em: http://people.sc.fsu.edu/~jburkardt/c_src/ball_grid/ball_grid.html. Acesso em: 17 abr. 2018.
- 116 SPHERE_GRID - grid on a sphere. 2017. Disponível em: http://people.sc.fsu.edu/~jburkardt/c_src/sphere_grid/sphere_grid.html. Acesso em: 17 abr. 2018.
- 117 BERNTSEN, J.; ESPELID, T. O.; GENZ, A. An adaptive algorithm for the approximate calculation of multiple integrals. **ACM Transactions on Mathematical Software**, v. 17, n. 4, p. 437–451, 1991.
- 118 GENZ, A.; MALIK, A. Remarks on algorithm 006: an adaptive algorithm for numerical integration over an n-dimensional rectangular region. **Journal of Computational and Applied Mathematics**, v. 6, n. 4, p. 295–302, 1980.
- 119 CUBATURE: multi-dimensional adaptive integration in C. 2017. Disponível em: <https://github.com/stevengj/cubature>. Acesso em: 17 abr. 2018.
- 120 LIU, H.; HEXEMER, A.; ZWART, P. H. The small angle scattering toolbox (SASTBX): an open-source software for biomolecular small-angle scattering. **Journal of Applied Crystallography**, v. 45, n. 3, p. 587–593, 2012.
- 121 NISIUS, B.; GOHLKE, H. Alignment-independent comparison of binding sites based on DrugScore potential fields encoded by 3D Zernike descriptors. **Journal of Chemical Information and Modeling**, v. 52, n. 9, p. 2339–2347, 2012.

-
- 122 VENKATRAMAN, V. et al. Protein-protein docking using region-based 3D Zernike descriptors. **BMC Bioinformatics**, v. 10, n. 1, p. 1–21, 2009.
- 123 GRANDISON, S.; ROBERTS, C.; MORRIS, R. J. The application of 3D Zernike moments for the description of “model-free” molecular structure, functional motion, and structural reliability. **Journal of Computational Biology**, v. 16, n. 3, p. 487–500, 2009.
- 124 SAEL, L. et al. Rapid comparison of properties on protein surface. **Proteins**, v. 73, n. 1, p. 1–10, 2008.
- 125 KIHARA, D. et al. Molecular surface representation using 3D Zernike descriptors for protein shape comparison and docking. **Current Protein and Peptide Science**, v. 12, n. 6, p. 520–530, 2011.
- 126 ZHU, X.; XIONG, Y.; KIHARA, D. Large-scale binding ligand prediction by improved patch-based method Patch-Surfer2.0. **Bioinformatics**, v. 31, n. 5, p. 707–713, 2014.
- 127 SIT, A. et al. An extension of 3D Zernike moments for shape description and retrieval of maps defined in rectangular solids. **Computational and Mathematical Biophysics**, v. 1, n. 1, p. 75–89, 2013.
- 128 LIU, H. et al. Computation of small-angle scattering profiles with three-dimensional Zernike polynomials. **Acta Crystallographica Section A**, v. 68, n. 2, p. 278–285, 2012.
- 129 CHOI, C. H. et al. Rapid and stable determination of rotation matrices between spherical harmonics by direct recursion. **The Journal of Chemical Physics**, v. 111, n. 19, p. 8825–8831, 1999.
- 130 KOVACS, J. A.; WRIGGERS, W. Fast rotational matching. **Acta Crystallographica Section D**, v. 58, n. 8, p. 1282–1286, 2002.