

UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE BIOLOGIA

JOÃO VICTOR DA SILVA GUERRA

PROSPECÇÃO E CARACTERIZAÇÃO DE CAVIDADES SUPRAMOLECULARES

PROSPECTION AND CHARACTERIZATION OF SUPRAMOLECULAR CAVITIES

CAMPINAS

2019

JOÃO VICTOR DA SILVA GUERRA

PROSPECÇÃO E CARACTERIZAÇÃO DE CAVIDADES SUPRAMOLECULARES

PROSPECTION AND CHARACTERIZATION OF SUPRAMOLECULAR CAVITIES

Dissertação apresentada ao Instituto de Biologia da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do Título de Mestre em Ciências, na área de Fármacos, Medicamentos e Insumos para Saúde.

Dissertation presented to the Institute of Biology of the University of Campinas in partial fulfillment of the requirements for the degree of Master of Science in Pharmaceuticals, Medicines and Health Supplies.

ESTE ARQUIVO DIGITAL CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO DEFENDIDA PELO ALUNO JOÃO VICTOR DA SILVA GUERRA E ORIENTADA PELO PROF. DR. PAULO SERGIO LOPES DE OLIVEIRA.

Orientador: DR. PAULO SERGIO LOPES DE OLIVEIRA

CAMPINAS

2019

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Biologia
Mara Janaina de Oliveira - CRB 8/6972

G937p

Guerra, João Victor da Silva, 1993-
Prospecção e caracterização de cavidades supramoleculares / João Victor da Silva Guerra. – Campinas, SP : [s.n.], 2019.

Orientador: Paulo Sergio Lopes de Oliveira.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. Software - Desenvolvimento. 2. Cavidades. 3. Caracterização. I. Oliveira, Paulo Sergio Lopes de. II. Universidade Estadual de Campinas. Instituto de Biologia. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Prospection and characterization of supramolecular cavities
Palavras-chave em inglês:
Computer software - Development
Cavities
Characterization
Área de concentração: Fármacos, Medicamentos e Insumos para Saúde
Titulação: Mestre em Ciências
Banca examinadora:
Paulo Sergio Lopes de Oliveira [Orientador]
Eduardo Xavier Silva Miqueles
Andre Luis Berteli Ambrosio
Data de defesa: 25-07-2019
Programa de Pós-Graduação: Biociências e Tecnologia de Produtos Bioativos

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-6800-4425
- Currículo Lattes do autor: http://lattes.cnpq.br/5809550322159439

Campinas, 25 de Julho de 2019.

COMISSÃO EXAMINADORA

Dr. Paulo Sergio Lopes de Oliveira

Dr. Eduardo Xavier Silva Miqueles

Dr. Andre Luis Berteli Ambrosio

Os membros da Comissão Examinadora acima assinaram a Ata de Defesa, que se encontra no processo de vida acadêmica do aluno.

DEDICATÓRIA

Dedico este trabalho aos meus pais, Mario Luiz da Silva Guerra e Roseli do Carmo Freitas da Silva, por todo apoio e amor, indispensáveis para a realização dos meus objetivos pessoais e profissionais.

AGRADECIMENTOS

Primeiramente, agradeço às pessoas que convivi durante estes anos e me apoiaram durante este projeto, sem os quais não seria possível a conclusão dessa importante etapa. Agradeço especialmente aos meus pais, Roseli do Carmo Freitas da Silva e Mario Luiz da Silva Guerra, que não pouparam esforços durante toda minha formação acadêmica e profissional. Agradeço também à minha namorada, Bruna Martins da Silva, que sempre me auxiliou e apoiou no desenvolvimento desse projeto.

Gostaria de agradecer ao meu orientador, Dr. Paulo Sergio Lopes de Oliveira, por ter me presenteado com um projeto que me identifiquei, a disponibilidade em auxiliar e guiar o desenvolvimento desta dissertação e principalmente a oportunidade de desenvolver meu trabalho dentro de um centro de pesquisa de excelência e propiciar a mudança de rumos na minha carreira acadêmica. Agradeço aos meus colegas do Laboratório de Biologia Computacional, José Geraldo de Carvalho Pereira, Helder Veras Filho, Juan Enrique Faya Castillo, Mariana Bortoletto Grizante e Rodrigo Vargas Honorato, que me apoiaram ao longo de momentos importantes, compartilhando ideias e sugestões para o desenvolvimento deste projeto. Palavras não podem medir a gratidão que tenho a todos que me ajudaram nessa etapa essencial da minha formação acadêmica.

Finalmente, agradeço a Pós-Graduação do Instituto de Biologia da Universidade Estadual de Campinas, aos membros do Programa de Biociências e Tecnologia de Produtos Bioativos, ao Laboratório Nacional de Biociências e ao Centro Nacional de Pesquisa em Energia e Materiais por terem viabilizado este trabalho. Também agradeço ao apoio financeiro da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo projeto de pesquisa regular (processo nº 2018/00629-0, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP)).

RESUMO

Interações proteicas desempenham funções vitais em processos biológicos. Para a execução dessas interações, cadeias proteicas se dobram orquestradamente de modo a criar sítios de ligação, que estão alocados em cavidades. Ao interagir com outras moléculas para formar complexos supramoleculares, as proteínas utilizam cavidades que não são facilmente identificáveis na forma não-ligada ou formam novas cavidades. A importância do seu estudo reside na compreensão, desenvolvimento e melhoramento de moléculas responsivas a complexos supramoleculares específicos. Nosso grupo de pesquisa do LNBio/CNPEM desenvolveu o programa KVFinder para a identificação de uma variedade de cavidades proteicas usando um sistema de dupla sonda. Entretanto, esta caracterização é somente topológica, calculando área, volume e forma das cavidades. Sabendo que as propriedades físico-químicas, espaciais e constitucionais são importantes para a determinação das funções e interações, é necessário mapear essas propriedades nas cavidades desses complexos. Desta forma, este trabalho introduz uma nova versão do programa KVFinder, parKVFinder, que manipula a escala de complexidade envolvida em complexos supramoleculares bioquimicamente e computacionalmente. Assim, a questão computacional foi abordada pela refatoração e paralelização das rotinas no parKVFinder, e a questão bioquímica, pelo desenvolvimento e aprimoramento de descritores moleculares mapeados sobre tais cavidades, abrangendo propriedades espaciais, constitucionais e físico-químicas. A caracterização espacial inclui volume, área, forma e profundidade das cavidades; a caracterização constitucional inclui os resíduos que formam as cavidades e suas características; e a caracterização físico-química inclui as escalas de hidrofobicidade e o potencial eletrostático mapeados nas cavidades. Por fim, o parKVFinder melhorou o desempenho de execução e a utilização dos recursos computacionais quando comparado ao KVFinder. A prova de conceito do parKVFinder foi realizada em uma proteína quinase A, apresentando uma descrição das propriedades das cavidades, que auxilia na compreensão das interações realizadas pela proteína.

ABSTRACT

Protein interactions play vital roles in biological processes. For the execution of these interactions, protein chains fold orchestrally in order to create binding sites, which are allocated in cavities. By interacting with other molecules to form supramolecular complexes, the proteins use cavities that are not easily identifiable in unbound form or create new cavities. The importance of its study lies in the understanding, development and improvement of molecules responsive to specific supramolecular complexes. Our LNBio/CNPEM research group has developed the KVFinder software for the identification of a variety of protein cavities using a dual probe system. However, this characterization is only topological, calculating area, volume and shape of the cavities. Knowing that the physicochemical, spatial and constitutional properties are important for the determination of functions and interactions, it is necessary to describe these properties in the cavities of these complexes. In this way, this work introduces a new version of the KVFinder program, parKVFinder, which manipulates the scale of complexity involved in supramolecular complexes biochemically and computationally. Thus, the computational question was addressed by the refactoring and parallelization of the routines in parKVFinder, and the biochemical issue, by the development and improvement of molecular descriptors mapped on such cavities, covering spatial, constitutional and physicochemical properties. Spatial characterization includes volume, area, shape and depth of cavities; constitutional characterization includes the residues that form the cavities and their characteristics; and the physicochemical characterization includes hydrophobicity scales and the electrostatic potential mapped in the cavities. Finally, parKVFinder improved execution performance and utilization of computational resources when compared to KVFinder. The proof of concept of parKVFinder was performed on a protein kinase A, presenting a description of the cavities properties, which helps in understanding the interactions performed by the protein.

LISTA DE ILUSTRAÇÕES

Figura 1: Representações das categorias de cavidades proteicas encontradas na estrutura tridimensional de proteínas. 19
Figura 2: Estrutura da proteína quinase A ligada ao peptídeo PKI e a adenosina. ...21
Figura 3: Representação da cavidade de acordo com propriedades físico-química dos resíduos.....22
Figura 4: Representação de uma grade tridimensional composta por voxels. 27
Figura 5: Representação esquemática da superfície molecular inserida na grade tridimensional. 28
Figura 6: Representação didática do funcionamento do KVFinder. 29
Figura 7: Prospecção de cavidades em uma toxina formadora de poros (PDB 4P24, rosa) para diferentes valores de sonda Probe Out. 30
Figura 8: Taxa de sucesso da detecção de sítios ativos pelo KVFinder. 31
Figura 9: Representação esquemática da execução de um conjunto de instruções através de computação serial e paralela.....34
Figura 10: Representação gráfica do *speedup* de um programa. 35
Figura 11: Fluxograma das rotinas empregadas no KVFinder.....39
Figura 12: Excerto de um arquivo PDB de cavidades biomoleculares. 43
Figura 13: Representação esquemática do funcionamento do parâmetro “distância de remoção”. 44
Figura 14: Arquivos padrões dos métodos de segregação do espaço na interface de linha de comando. 46
Figura 15: Classes de voxels de superfície.....47
Figura 16: Representação dos espaços de busca de pontos de cavidades por átomos da estrutura biomolecular. 49
Figura 17: Classes de aminoácidos descrito em Lehninger, Nelson e Cox (1995)....50
Figura 18: Dicionário de escalas de hidrofobicidade.....52
Figura 19: Distribuição de tipos moleculares no repositório RCSB PDB. 54
Figura 20: Distribuição de tipos enzimáticos no repositório RCSB PDB.....55
Figura 21: Distribuição de massa molecular das entidades distintas no repositório RCSB PDB. 56
Figura 22: Estatísticas do conjunto de testes kv1000. 57

Figura 23: Estrutura da grande subunidade do *Staphylococcus aureus* em complexo com lincomicina. 58
Figura 24: Fluxograma das rotinas empregadas no programa KVFinder atualizado e otimizado. 62
Figura 25: Desempenho computacional das versões do KVFinder.. 63
Figura 26: Desempenho computacional do programa parKVFinder. 66
Figura 27: Representação esquemática de pontos vizinhos e pontos adjacentes. ...68
Figura 28: Comparação dos filtros de ponto de superfície. 68
Figura 29: Investigação dos efeitos do parâmetro de distância de remoção. 70
Figura 30: Investigação do novo parâmetro de resolução. 72
Figura 31: Menu de ajuda da interface de linha de comando do programa parKVFinder. 74
Figura 32: Conjunto de sólidos geométricos ocos. 77
Figura 33: Resultados da metodologia de determinação da profundidade das cavidades biomoleculares. 79
Figura 34: Determinação da profundidade dos sítios de ligação da quinase A. 80
Figura 35: Determinação dos resíduos formadores do sítio de ligação da adenosina da proteína quinase A. 81
Figura 36: Composição, características e contagem dos resíduos formadores da cavidade do sítio de ligação da adenosina. 82
Figura 37: Resultados da metodologia das escalas de hidrofobicidade nas cavidades proteicas. 84
Figura 38: Determinação da hidropatia do sítio de ligação da adenosina da quinase A. 85
Figura 39: Determinação do potencial eletrostático dos sítios de ligação da quinase A. 87

LISTA DE TABELAS

Tabela 1: Exemplos de métodos para detecção de sítios de ligação. 26
Tabela 2: Informações das funções do programa KVFinder pela ferramenta *Valgrind*. 59
Tabela 3: Volume de voxel definido por cada opção de resolução no programa parKVFinder. 71
Tabela 4: Área superficial das cavidades do conjunto de sólidos geométricos ocos.77

LISTA DE ABREVIATURAS E SIGLAS

3D	Tridimensional
API	Interface de programação de aplicações
CNPEM	Centro Nacional de Pesquisa em Energia e Materiais
DFS	Busca em profundidade
IPL	Interação proteína-ligante
IPP	Interação proteína-proteína
JSON	*JavaScript Object Notation*
LBC	Laboratório de Biologia Computacional
LNBio	Laboratório Nacional de Biociências
logP	Coeficiente de partição P
PBE	Equação de Poisson-Boltzmann
PDB	Banco de Dados de Proteínas
PDBe	Banco de Dados de Proteínas da Europa
PDBj	Banco de Dados de Proteínas do Japão
PKI	Peptídeo inibidor de quinase
parKVFinder	*Parallel* KVFinder
RCSB PDB	Banco de Dados de Proteínas da *Research Collaboratory for Structural Bioinformatics*
TOML	*Tom's Obvious, Minimal Language*
YAML	*YAML Ain't Markup Language*

LISTA DE SÍMBOLOS

$\widehat{A_i}$	Área superficial estimada da cavidade i
d_p	Diâmetro da sonda *Probe Out*
E_p	Eficiência do programa para p processos
f_{np}	Fração não-paralelizável do programa
f_p	Fração paralelizável do programa
h	Espaçamento de grade
m	Unidades de grade no eixo X
n	Unidades de grade no eixo Y
N_p	Número de processadores disponíveis
$N_{S,i}$	Número de voxels de superfície pertencentes a cavidade i
$N_{S,j,i}$	Número de voxels de superfície da classe j pertencentes a cavidade i
$N_{V,i}$	Número de voxels pertencentes a cavidade i
o	Unidades de grade no eixo Z
p	Número de processos
P_i	Ponto i de coordenadas x_i, y_i e z_i
S	*Speedup* entre dois programas
S_A	*Speedup* estimado pela lei de Amdahl
S_G	*Speedup* estimado pela lei de Gustafson
S_p	*Speedup* do programa para p processos
t_i	Tempo computacional de execução do programa i
$\widehat{V_i}$	Volume estimado da cavidade i
W_j	Peso da classe j de voxels de superfície
x_i	Coordenada ortogonal no eixo X em ångström
y_i	Coordenada ortogonal no eixo Y em ångström
z_i	Coordenada ortogonal no eixo Z em ångström

SUMÁRIO

1 INTRODUÇÃO 16

1.1 Compreensão do processo de reconhecimento molecular 16

1.2 Prospecção de cavidades proteicas 18

1.2.1 Classificação de cavidades 19

1.3 Caracterização de cavidades biomoleculares 20

1.3.1 Propriedades espaciais 20

1.3.2 Propriedades físico-químicas 23

1.4 Estado da arte 24

1.5 Programa KVFinder 27

1.5.1 Metodologia para prospecção de cavidades biomoleculares 28

1.5.2 Metodologia para estimativa de volume e área superficial 29

1.5.3 Desafios na detecção de cavidades superficiais 31

1.5.4 Limitações computacionais 32

1.6 Escala supramolecular 32

1.7 Computação paralela 33

2 JUSTIFICATIVA 36

3 OBJETIVOS 37

3.1 Geral 37

3.2 Específicos 37

4 METODOLOGIA 37

4.1 Atualização e otimização 38

4.2 Implementação de computação paralela 41

4.3 Melhorias incrementais no parKVFinder 42

4.3.1 Definição dos pontos de superfície 43

4.3.2 Implementação do parâmetro "distância de remoção" 43

4.3.3 Implementação de método indireto do espaçamento de grade 45

4.3.4 Desenvolvimento de interface de linha de comando 45

4.4 Desenvolvimento de conjunto de testes 46

4.5 Implementação de descritores de propriedades 46

4.5.1 Descritores de propriedades espaciais 46

4.5.2 Descritores de propriedades constitucionais 48

4.5.3 Descritores de propriedades físico-químicas 51

5 RESULTADOS E DISCUSSÃO 54

5.1 Desenvolvimento do conjunto de testes 54

5.2 Atualização, otimização e computação paralela das rotinas 58

5.2.1 Análise do desempenho computacional da atualização e otimização 62

5.2.2 Análise do desempenho computacional da paralelização 64

5.3 Melhorias incrementais no parKVFinder 67

5.3.1 Definição dos pontos de superfície 67

5.3.2 Implementação do parâmetro "distância de remoção" 69

5.3.3 Implementação do parâmetro "resolução" 71

5.3.4 Desenvolvimento da interface de linha de comando 73

5.4 Descritores de propriedades 76

5.4.1 Descritores de propriedades espaciais 76

5.4.2 Descritores de propriedades constitucionais 80

5.4.3 Descritores de propriedades físico-químicas 83

6 CONCLUSÕES E PERSPECTIVAS 89

7 REFERÊNCIAS 91

ANEXO A: EXEMPLOS DE ARQUIVOS DE CONFIGURAÇÃO 100

ANEXO B: ARQUIVOS DO PROGRAMA KVFINDER 102

ANEXO C: DIRETIVAS DE COMPILAÇÃO DO OPENMP 106

ANEXO D: HISTOGRAMAS CONSTITUCIONAIS DO PARKVFINDER 107

ANEXO E: DECLARAÇÕES 109

1 INTRODUÇÃO

1.1 Compreensão do processo de reconhecimento molecular

A química medicinal é um campo de pesquisa que investiga compostos orgânicos bioativos, denominados biosondas, usando conhecimentos interdisciplinares em biologia, química e medicina. As biosondas têm sido empregadas para investigar processos celulares importantes através da elucidação de mecanismos relacionados, sendo potenciais pistas para o desenvolvimento de novas terapias e drogas (OSADA, 2009). Por outro lado, uma droga precisa interagir com uma molécula alvo, geralmente uma proteína, para executar uma função fisiológica específica (SOTRIFFER; KLEBE, 2002). Experiências mostram que as interações proteína-ligante (IPLs) em sítios de ligação dependem das propriedades físico-químicas e espaciais (SOTRIFFER; KLEBE, 2002; OSADA, 2009; OLIVEIRA et al., 2014). As IPLs são altamente específicas e variáveis através de domínios proteicos e classes de ligantes, restringindo assim a um pequeno número de ligantes a interação eficiente com uma dada proteína (HENRICH et al., 2010; OLIVEIRA et al., 2014). Tais especificidades dependem de um microambiente que atenda a restrições geométricas do receptor, como complementaridade espacial e grandes interfaces de interação (SOTRIFFER; KLEBE, 2002; HENRICH et al., 2010). Com base nisso, a investigação das propriedades do sítio de ligação é um ponto de partida da química medicinal.

Dada a importância dos mecanismos das IPLs, o papel fundamental para a compreensão da ação do ligante e, consequentemente, a identificação e caracterização de sítios de ligação funcionalmente relevantes são um ponto de partida importante para o desenho racional de fármacos (SOTRIFFER; KLEBE, 2002; OSADA, 2009). Sendo assim, a informação estrutural atômica tridimensional (3D) da proteína é uma fonte fundamental para compreender os processos biológicos e explorar os mecanismos de ação do fármaco (ROSE et al., 2017). Um banco de dados global, o Banco de Dados de Proteínas (*Protein Data Bank;* PDB), para estruturas 3D de macromoléculas biológicas (proteínas, DNA e RNA) determinadas experimentalmente, auxilia na identificação de sítios de ligação funcionalmente relevantes, fornecendo acesso à estrutura 3D resolvida por cristalografia de difração de raios X, espectroscopia de ressonância magnética nuclear e/ou crio-microscopia eletrônica (SOTRIFFER; KLEBE, 2002; DABERDAKU; FERRARI, 2016; ROSE et al.,

2017). O PDB inclui três repositórios regionais: Banco de Dados de Proteínas da *Research Collaboratory for Structural Bioinformatics* (RCSB PDB; http://rcsb.org) nos EUA (ROSE et al., 2017), Banco de Dados de Proteínas do Japão (PDBj; http://pdbj.org) no Japão (KINJO et al., 2012) e Banco de Dados de Proteínas da Europa (PDBe; http://pdbe.org) na Europa (VELANKAR et al., 2016).

Uma vez que a estrutura atômica 3D de uma proteína esteja disponível, a estratégia para projetar racionalmente um ligante depende de informações adicionais sobre os alvos biológicos relevantes (SOTRIFFER; KLEBE, 2002). Normalmente, existem três situações distintas:

I. A localização do sítio de ligação é desconhecida

Normalmente, pelo menos uma função da proteína alvo é conhecida; porém, pode não estar totalmente elucidada em termos bioquímicos e estruturais. Esta situação requer um método para identificar o sítio de ligação, em que um ligante pode ser planejado para inibir ou promover a função da proteína alvo. No entanto, nos casos em que a função biológica não foi atribuída, podem ser necessários métodos para inferir a função da proteína com base na estrutura atômica (SOTRIFFER; KLEBE, 2002). Uma vez que a função da proteína é comumente ligada à interação, elucidar as características funcionais e identificar os sítios de ligação estão relacionados.

II. A localização aproximada do sítio de ligação é conhecida, mas nenhuma informação sobre as propriedades do sítio e interações do ligante está disponível

Esta situação representa aproximadamente 30% das estruturas de proteínas armazenadas no PDB (NUÑEZ-VIVANCO et al., 2016), o que requer uma análise minuciosa da estrutura da proteína na área do sítio de ligação, apresentando regiões de interação favoráveis onde certos grupos funcionais podem se ligar. Os métodos fornecem um mapa funcional do sítio de ligação que orienta o posicionamento ou desenvolvimento de ligantes potenciais ou cria um modelo de farmacóforo para pesquisa em banco de dados.

III. A localização do sítio de ligação e a interação do ligante são conhecidas

Dado um sítio de ligação bem conhecido, os métodos para determinar o mapa funcional ou energético do sítio de ligação por acoplamento molecular (*docking*) e triagem virtual (*virtual screening*) são ferramentas úteis. Estes mapas funcionais ou energéticos são abordagens apropriadas para orientar o posicionamento de potenciais ligantes ou o desenvolvimento de ligantes adequados.

Abordagens computacionais para identificar sítios de ligação e uma coleção de propriedades de sítios de ligação serão introduzidas a seguir. Os métodos computacionais empregam a riqueza de informações estruturais de complexos proteína-ligante e proteínas não-ligadas para identificar potenciais sítios de ligação. O fenômeno de ligação do complexo proteína-ligante é governado pelas propriedades espaciais e físico-químicas do sítio de ligação, produzindo uma alta complementaridade entre a proteína e o ligante (SOTRIFFER; KLEBE, 2002; OSADA, 2009; HENRICH et al., 2010). Juntos, eles ajudam a esclarecer deficiências na compreensão do reconhecimento molecular em sistemas biológicos.

1.2 Prospecção de cavidades proteicas

Os sítios de ligação geralmente estão localizados em uma ampla variedade de cavidades proteicas, como bolsões, invaginações, fendas e canais (SOTRIFFER; KLEBE, 2002). A prospecção e caracterização dos sítios de ligação são essenciais para a compreensão das complexas interações das proteínas e suas respectivas funções. As características topológicas e as propriedades físico-químicas apresentadas na interface de contato, como forma, volume, área, distribuição de cargas, hidrofobicidade, solvatação e tipo de resíduo constituinte, ditam as funções e interações das proteínas (HUBBARD; ARGOS, 1994; BOHACEK; MCMARTIN, 1997; OLIVEIRA et al., 2014; DABERDAKU; FERRARI, 2016). A identificação e dimensionamento das cavidades são os primeiros passos para projetar um ligante baseado na estrutura proteica (LIANG; WOODWARD; EDELSBRUNNER, 1998). Além disso, as características estruturais e físico-químicas da proteína são valiosas na descoberta e otimização de fármacos (DABERDAKU; FERRARI, 2016).

O avanço da proteômica na última década trouxe à tona um repositório de possíveis alvos para fármacos, na forma de dados sobre interações supramoleculares ou interactoma. Proteínas participam em papéis essenciais de diversos processos celulares, como, por exemplo, transdução de sinal, adesão celular, proliferação celular e apoptose (SCOTT et al., 2013; PIETERS et al., 2016; JANA et al., 2017), principalmente através da interação, desde de pequenos íons, como fosfato e ferro, até supramoléculas, como ribossomos e nucleossomos (OLIVEIRA et al., 2014; PIETERS et al., 2016). Muitas patologias, como câncer, doenças metabólicas e patogênicas, estão associadas a desregulação dessas interações (SCOTT et al., 2013). A cada dia se torna mais evidente a necessidade de identificar regiões na área

de interação que podem ser utilizadas como alvo para novos fármacos.

1.2.1 Classificação de cavidades

As proteínas apresentam uma alta complexidade estrutural, sendo repletas de espaços vazios e espaços superficiais, conhecidos como cavidades proteicas (OLIVEIRA, 2011). Essas diferentes formas topológicas apresentam uma grande variação em suas propriedades espaciais e físico-químicas, criando um microambiente único capaz de interagir especificamente com outras moléculas (LIANG; WOODWARD; EDELSBRUNNER, 1998).

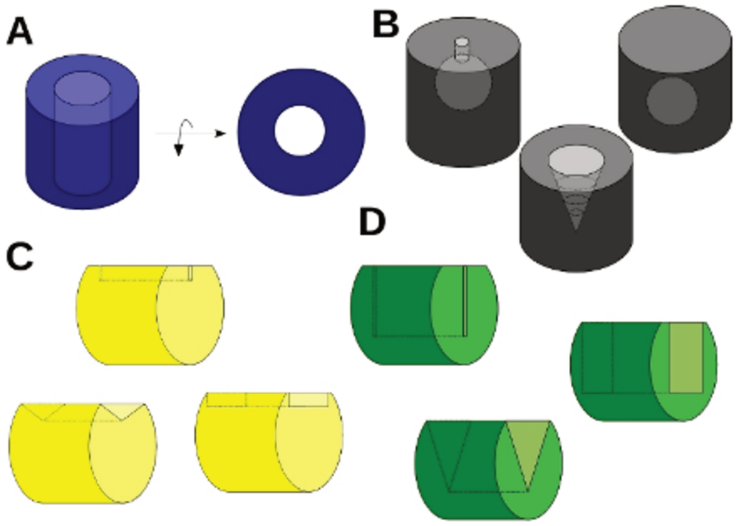


Figura 1: Representações das categorias de cavidades proteicas encontradas na estrutura tridimensional de proteínas. (A) Túneis. **(B)** Vazios enclausurados (direita), bolsões (centro) e invaginações (esquerda). **(C)** Elementos superficiais. **(D)** Vales proteicos.

As cavidades proteicas podem ser classificadas de acordo com a sua topologia em: túneis, vazios enclausurados, bolsões, invaginações, elementos superficiais e vales proteicos (Figura 1). Os túneis (Figura 1A) são espaços vazios que se estendem ao longo de um dos eixos da estrutura proteica, possuindo mais de um ponto de acesso. Os vazios enclausurados, bolsões e invaginações (Figura 1B) são espaços vazios compreendidos no interior de uma estrutura proteica. Os vazios enclausurados são cavidades que não possuem pontos de acesso. Os bolsões possuem uma área de acesso com área relativamente maior que o tamanho da cavidade. As invaginações possuem uma área de acesso com área relativamente menor que o tamanho da cavidade. Os elementos superficiais (Figura 1C) são depressões que se estendem pela superfície da estrutura sem apresentar grande profundidade, como fendas, fissuras e gretas, sendo amplamente encontrados na região da interface da interação

proteína-proteína (IPP). Os vales proteicos (Figura 1D) se estendem ao longo de um dos eixos da proteína e são altamente expostos ao meio externo, além de apresentar ponto de acesso com alto grau de abertura (OLIVEIRA, 2011).

De acordo com Hubbard e Argos (1994), as cavidades proteicas também podem ser classificadas espacialmente em: intradomínio ou interdomínio, inter-subunidade ou interfacial de subunidade, e superficiais acessíveis ou superficiais oclusas. Cada tipo de classificação espacial apresenta propriedades estruturais, físico-químicas e constitucionais diferentes entre si, sendo que essas propriedades afetam sua estabilidade e suas funções.

1.3 Caracterização de cavidades biomoleculares

Como as diferentes interações realizadas em uma cavidade são delimitadas por suas propriedades, essas interações determinam a função biológica da biomolécula (SOTRIFFER; KLEBE, 2002). A partir disso, as propriedades espaciais e físico-químicas em uma cavidade biomolecular podem ser utilizadas para identificar potenciais sítios de ligação relevantes às interações estudadas.

O fenômeno de IPL e IPP em sítios de ligação é uma consequência das propriedades das cavidades proteicas que governam o reconhecimento molecular do ligante, variando de uma pequena molécula a uma supramolécula (SOTRIFFER; KLEBE, 2002). A base do reconhecimento molecular depende da complementaridade espacial e físico-química entre a proteína alvo e seu ligante (HENRICH et al., 2010; DABERDAKU; FERRARI, 2016). Para isso, o sítio de ligação determina restrições espaciais e físico-químicas a serem atendidas por seus ligantes putativos (SOTRIFFER; KLEBE, 2002), sendo que cada interação com ligante pode desencadear uma função biológica diferente. Portanto, a descrição de potenciais sítios de ligação em aspectos espaciais e físico-químicos é essencial para o desenho racional e otimização de fármacos, e também para a avaliação da *drogabilidade* de sítios de ligação (HUBBARG; ARGOS, 1994; LIANG; WOODWARD; EDELSBRUNNER, 1998; DABERDAKU; FERRARI, 2016).

1.3.1 Propriedades espaciais

As características espaciais de cavidades biomoleculares estabelecem restrições quanto à geometria dos ligantes que podem interagir eficientemente nelas. Alta afinidade entre o sítio de ligação e seus potenciais ligantes depende de interfaces

de interação suficientemente grandes, em outras palavras, extensa área superficial na cavidade biomolecular. A especificidade do sítio é imposta pelas restrições geométricas tais como tamanho, forma e extensão do enterramento das cavidades (LASKOWSKI et al., 1996; LIANG; WOODWARD; EDELSBRUNNER, 1998; DABERDAKU; FERRARI, 2016). Notavelmente, o volume da cavidade deve acomodar o potencial ligante (STANK et al., 2016). A complementaridade de forma entre a proteína e o ligante é um fator determinante do processo de ligação, comumente os pequenos ligantes se ligam em sítios de ligação côncavos na superfície da proteína (HENRICH et al., 2010). Em estudos comparativos, foi demonstrado que, para algumas famílias enzimáticas, o sítio ativo é geralmente caracterizado por um vale proteico de grande extensão e profundidade (LASKOWSKI et al., 1996). Essas características podem ser observadas em proteínas quinases, fundamentais para o crescimento celular e transdução de sinal, onde o sítio ativo responsável pela fosforilação do substrato é localizado em um vale proteico (Figura 2). Descrições desse tipo podem auxiliar na determinação da funcionalidade da cavidade proteica.

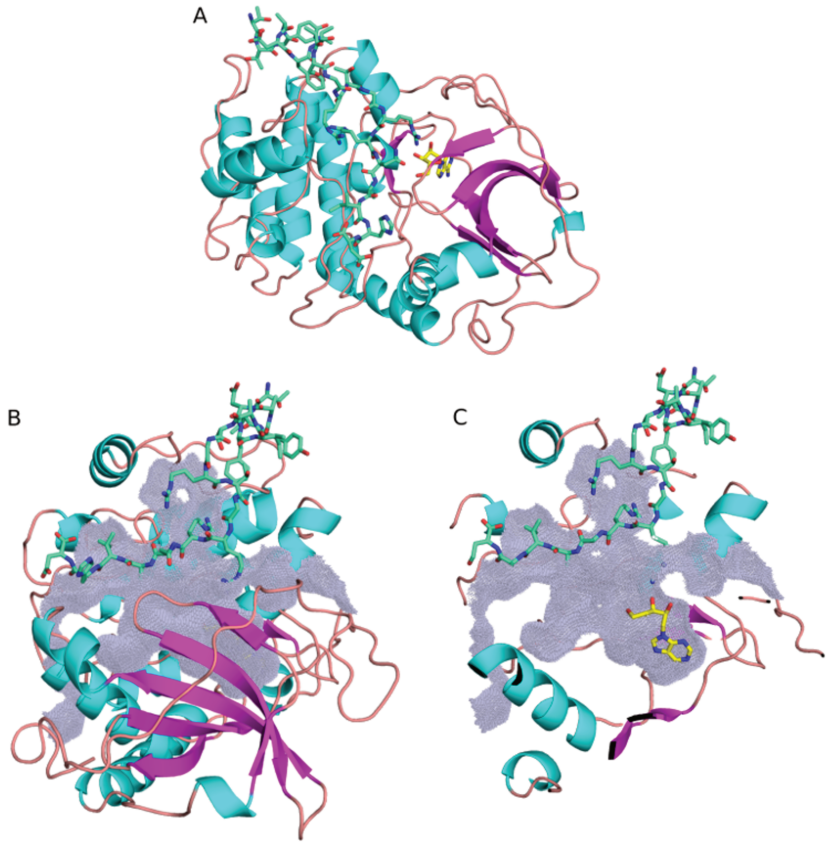


Figura 2: Estrutura da proteína quinase A ligada ao peptídeo PKI e a adenosina. (A) Proteína quinase A representada em *cartoon* e substrato PKI em *sticks*. **(B)** Resultado da detecção do vale catalítico pelo KVFinder. **(C)** Visão interna do vale catalítico evidenciando o sítio do ATP (amarelo).

A inferência de funcionalidade das cavidades proteicas tem sido feita através da comparação com proteínas homólogas e a função de sítios ativos é determinada por homologia e similaridade sequencial (JUNCKER et al., 2009). O papel funcional de uma proteína também pode ser determinado por meio de sua estrutura 3D (dobramento espacial) que apresenta um grau de conservação superior à sequencial. Abordagens atuais buscam comparar o dobramento proteico por meio de banco de dados como o PDB (BERMAN, 2000) e o CATH (SILLITOE et al., 2014). Esses bancos disponibilizam informações essenciais para investigar a funcionalidade e localização de sítios de ligação. Além dos padrões globais de identidade sequencial e homologia estrutural, os padrões estruturais locais também são importantes para determinação da função. Um exemplo desse padrão estrutural local e conservado são as tríades catalíticas de serino-proteases (DODSON; WLODAWER, 1998). Em enzimas, essas configurações específicas assumem arranjos espaciais preferenciais para desempenharem etapas elementares da reação catalítica e ditarem o processo de reconhecimento molecular dos seus ligantes pelas cavidades ativas (THORNTON et al., 2000; SOTRIFFER; KLEBE, 2002; HENRICH et al., 2010). A partir desse racional, é interessante descrever uma cavidade proteica de acordo com a localização de seus diferentes tipos de resíduos; doadores de hidrogênios, receptores de elétrons, contatos hidrofóbicos e aromáticos (Figura 3).

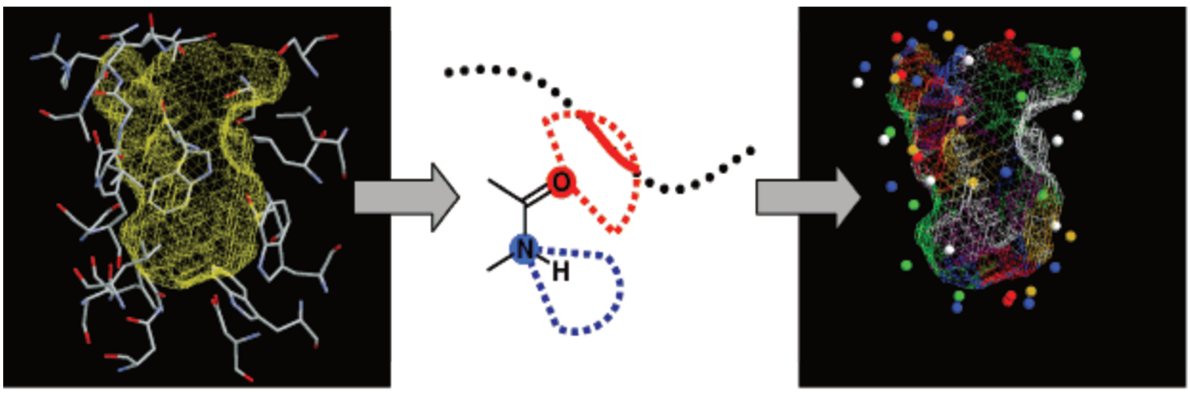


Figura 3: Representação da cavidade de acordo com propriedades físico-química dos resíduos. Doador de hidrogênio (azul), aceptor de hidrogênio (vermelho), doador/aceptor (verde), acesso à aminoácido hidrofóbico (branco) e aromático (laranja) (SCHIMITT; HENDLICH; KLEBE, 2001).

Os sítios de ligação possuem suas sequências de aminoácidos conservadas dentro das famílias de proteínas, que podem fornecer informações funcionais (HENRICH et al., 2010). A composição de aminoácidos pode caracterizar sítios de

ligação enzimáticos e não-enzimáticos, que apresentaram maior frequência de resíduos específicos em diferentes locais e tipos de proteínas (BARTLETT et al., 2002; SOGA et al., 2007; CARLSON et al., 2008). Nos sítios catalíticos, os resíduos carregados foram encontrados com mais frequência do que os polares e os hidrofóbicos em sua composição (BARTLETT et al., 2002). Além disso, diferentes frequências de aminoácidos foram observadas nos sítios de ligação em comparação com outros sítios da proteína (SOGA et al., 2007). A afinidade dos sítios de ligação é governada pela composição de aminoácidos, apresentando diferentes composições entre sítios de ligação enzimáticos e não-enzimáticos (CARLSON et al., 2008). Portanto, a descrição da localização e características físico-químicas dos resíduos estruturais no sítio de ligação permite a classificação das cavidades de acordo com suas possíveis funções e direciona de maneira assertiva o desenho racional de fármacos.

1.3.2 Propriedades físico-químicas

Possíveis IPLs não dependem apenas da complementaridade geométrica entre o ligante e o sítio de ligação, mas também dependem da complementaridade físico-química entre eles. A contribuição das interações de van der Waals, hidrofóbica, eletrostática, ligação de hidrogênio e solvatação deve resultar em um ambiente energeticamente favorável para o processo de ligação (MORRIS et al., 2009; HENRICH et al., 2010). Aqui, consideramos algumas importantes propriedades físico-químicas dos sítios de ligação.

O potencial eletrostático é usualmente calculado pela equação de Poisson-Boltzmann (PBE) (HONIG; NICHOLLS, 1995), aplicada por diferentes abordagens numéricas (XIE; YING; XIE, 2017; HARRIS et al., 2017). Tendo o modelo de solvente implícito como o mais popular, uma ampla gama de programas foram desenvolvidas, como GRASP (NICHOLLS; SHARP; HONIG, 1991), UHBD (MADURA et al., 1995), PBEQ (JO et al., 2008), APBS (UNNI et al., 2011), DelPhi (SMITH et al., 2012) e PSBA (WANG et al., 2013). A partir deste potencial eletrostático estimado, informações sobre possíveis pontos de ancoragem de ligantes, energia livre de interação, estabilidade das biomoléculas e forças atômicas médias podem ser obtidas.

A hidrofobicidade de uma molécula é descrita pelo seu coeficiente de partição P, representado por logP, entre as fases de dois solventes imiscíveis sob condições de equilíbrio (HENRICH et al., 2010). O logP mais comum usado na ciência

farmacêutica é o coeficiente de partição entre as fases de octanol e água, porque descreve a diferença entre o plasma polar e as membranas celulares lipofílicas (OBERHAUSER; NURISSO; CARRUPT, 2014). A hidrofobicidade desempenha um papel importante nas características cinéticas e dinâmicas da ação de fármacos (MANNHOLD et al., 2009). A intensidade da interação hidrofóbica é governada pela forma do sítio de ligação e pela área exposta dos resíduos (HENRICH et al., 2010). No entanto, a previsão de interações hidrofóbicas para aminoácidos é complicada e diferentes abordagens geram diferentes escalas de hidrofobicidade (HEIDEN; MOECKEL; BRICKMANN, 1993). Os métodos atuais para prever os coeficientes de partição são baseados em subestruturas ou abordagens de propriedades, por exemplo, HINT (KELLOGG; SEMUS; ABRAHAM, 1991), ALOGPS (TETKO; TANCHUK, 2002), KLOGP (ZHU et al., 2005) e MLP (OBERHAUSER; NURISSO; CARRUPT, 2014). Abordagens baseadas em subestruturas reduzem as moléculas em fragmentos ou átomos e somam essas contribuições de subestruturas para estimar o logP final. Abordagens baseadas em propriedades aplicam descritores moleculares e topológicos de toda a molécula, usando abordagens empíricas para representação da estrutura 3D (MANNHOLD et al., 2009).

1.4 Estado da arte

O desenvolvimento de métodos computacionais para prospecção e caracterização de sítios de ligação em proteínas são essenciais para aprofundar os conhecimentos acerca da função de uma determinada proteína e para descoberta e melhoramento de fármacos (LIANG; WOODWARD; EDELSBRUNNER, 1998). Os métodos baseados em dados estruturais são mais informativos do que os métodos baseados em sequências de aminoácidos, sendo que as estruturas proteicas 3D são mais conservadas do que as sequências (NUÑEZ-VIVANCO et al., 2016). Baseados nestes conhecimentos, métodos computacionais foram desenvolvidos para identificar potenciais sítios de ligação (OLIVEIRA et al., 2014; NUÑEZ-VIVANCO et al., 2016; DABERDAKU; FERRARI, 2016). Os algoritmos publicados se dividem em três categorias principais de acordo com a metodologia utilizada: geométricos, energéticos e evolutivos (OLIVEIRA et al., 2014; DABERDAKU; FERRARI, 2016).

Os métodos geométricos se baseiam na análise da superfície molecular e compõe a maioria dos algoritmos disponíveis. As representações da superfície proteica variam em diferentes métodos, sendo baseados em técnicas de voxels (pixel

volumétrico), também conhecidas como representações de grade tridimensional, esferas e tesselação, também conhecidas como representações de malha triangular (OLIVEIRA et al., 2014; DABERDAKU; FERRARI, 2016). Por outro lado, os métodos energéticos identificam sítios de ligação a partir da análise das forças intermoleculares ou a energia de interação entre a proteína alvo e uma sonda, normalmente um grupo químico. Por fim, ferramentas que utilizam métodos evolutivos se baseiam na busca por resíduos conservados, alinhamentos de sequências e informações de perfis de sítios de ligação conhecidos. Combinações das abordagens mencionadas acima também foram publicadas, melhorando a taxa de predição de sítios de ligação. Exemplos e classificação dos métodos para detecção de sítios de ligação estão apresentados na Tabela 1.

Métodos baseados em princípios geométricos apresentam algumas vantagens quando comparados a métodos energéticos e evolutivos. Os métodos baseados em geometria não dependem de conhecimento prévio quando comparados a abordagens evolutivas, sendo independentes de informações de sequência ou de bancos de dados de sítios de ligação ativos. Em contraste com os métodos baseados em energia, os métodos geométricos são uma abordagem mais direta, que são independentes das funções de pontuação e da parametrização de campo de força. Além disso, métodos puramente geométricos para identificação de sítios de ligação são eficientes para identificar todas as cavidades de uma dada proteína alvo. Contudo, o problema consiste em reconhecer quais cavidades da proteína são sítios de ligação funcionalmente relevantes (SOTRIFFER; KLEBE, 2002; HENRICH et al., 2010). Embora não exista uma abordagem simples e eficaz, a caracterização dos sítios de ligação em termos de propriedades espaciais e físico-químicas pode conduzir à identificação de sítios de ligação funcionalmente relevantes.

Em métodos baseados em geometria, existem diferentes abordagens para determinação e representação de superfícies moleculares e suas propriedades, tais como forma-alfa, grade tridimensional, funções gaussianas tridimensionais, imagens de spin e tesselação, diferindo em eficiência e simplicidade (DABERDAKU; FERRARI, 2016). Métodos *in silico* para prever funções, propriedades e interações exigem representações adequadas da superfície molecular. Como a complementaridade espacial e físico-química entre a proteína e o ligante é necessária, a escolha de uma representação de superfície é um passo importante para compreender seus papéis nos processos biológicos (HENRICH et al., 2010; DABERDAKU; FERRARI, 2016;

STANK et al., 2016). Dentre as diversas representações de superfície molecular, a grade tridimensional composta por voxels é a mais simples e apropriada para a representação de múltiplas propriedades em várias condições, pois cada voxel da grade tridimensional pode acumular diferentes informações.

Tabela 1: Exemplos de métodos para detecção de sítios de ligação.

Programa	Classificação	Referência
AutoDock	Energético	Morris et al. (2009)
CAST	Geométrico	Liang, Woodward e Edelsbrunner (1998)
Cavbase	Geométrico	Kuhn et al. (2006)
CAVER	Evolutivo	Petrek et al. (2006)
CavitySearch	Geométrico	Ho e Marshall (1990)
ConSurf	Evolutivo	Armon, Graur e Ben-Tal (2001)
CS-Map	Energético	Bradford e Westhead (2005)
DogSite	Geométrico	Volkamer et al. (2010)
DrugSite	Energético	An, Totrov e Abagyan (2004)
FINDSITE	Combinado	Brylinski e Skolnick (2008)
Fpocket	Geométrico	Guilloux, Schmidtke e Tuffery (2009)
GRID	Energético	Goodford (1985)
KVFinder	Geométrico	Oliveira et al. (2014)
LIGSITE	Geométrico	Hendlich, Rippman e Barnickel (1997)
LigSiteCSC	Combinado	Huang e Schoroeder (2006)
MetaPocket	Combinado	Huang (2009)
PASS	Geométrico	Brady e Stouten (2000)
PHECON	Geométrico	Kawabata e Go (2007)
POCASA	Geométrico	Yu et al. (2010)
POCKET	Geométrico	Levitt e Banaszak (1992)
PocketFinder	Energético	An, Totrov e Abagyan (2005)
PocketPicker	Geométrico	Weisel, Proschak e Schneider (2007)
QsiteFinder	Energético	Laurie e Jackson (2005)
Rate4Site	Evolutivos	Pupko et al. (2002)
SCREEN	Geométrico	Nayal e Honig (2006)
SiteHound	Energético	Ghersi e Sanchez (2009)
SiteMap	Combinado	Halgren (2007)
SURFNET	Geométrico	Laskowski (1995)
TRAPP	Geométrico	Kokh et al. (2013)
VOIDOO	Geométrico	Kleywegt e Jones (1994)

1.5 Programa KVFinder

O programa KVFinder (OLIVEIRA et al., 2014) foi desenvolvido pelo Laboratório de Biologia Computacional (LBC) do Laboratório Nacional de Biociências (LNBio) no Centro Nacional de Pesquisa em Energia e Materiais (CNPEM) para prospecção e caracterização espacial de qualquer tipo de cavidade biomolecular. Esse programa, implementado em ANSI C, tem código livre e aberto, e é distribuído gratuitamente à comunidade científica. A ferramenta é disponibilizada ao lado de um *plugin* gráfico, implementado em Python, para programa de visualização PyMOL (SCHRÖDINGER, LLC, 2015) com a finalidade de garantir a máxima usabilidade do KVFinder de usuários básicos a especialistas da área. Em Oliveira et al. (2014), o programa foi validado, como uma metodologia de detecção de cavidades, e seus resultados comparados com outros programas.

A prospecção de cavidades biomoleculares é realizada com uma abordagem geométrica baseada em grade tridimensional subdividida em voxels. O voxel representa um ponto discreto de dados em uma grade regular no espaço tridimensional, sendo que cada ponto pode conter mais de uma informação a fim de representar diferentes propriedades em uma certa porção de espaço de maneira simples e efetiva (Figura 4).

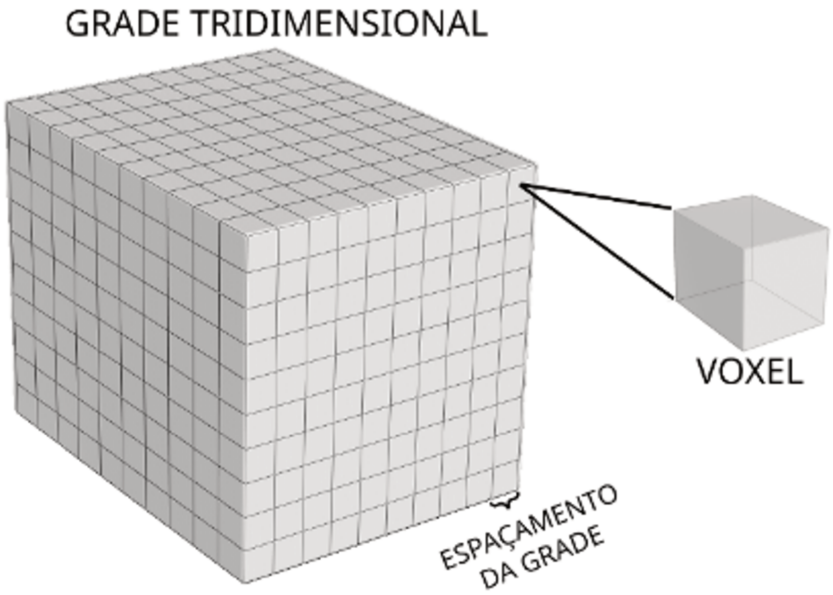


Figura 4: Representação de uma grade tridimensional composta por voxels.

O KVFinder apresenta algumas capacidades distintas, como a segmentação do espaço de busca e um conjunto customizável de parâmetros, que permite a

representação de cavidades em alta resolução. Esses parâmetros definidos pelo usuário são projetados para resolver algumas das principais falhas de métodos geométricos, como espaçamento de grade, definição dos limites e teto das cavidades e segmentação de espaço.

1.5.1 Metodologia para prospecção de cavidades biomoleculares

O primeiro desafio para qualquer método de prospecção de cavidades biomoleculares é a própria definição da cavidade. Uma definição formal não foi criada para cavidades biomoleculares, o que leva a uma discrepância na maneira como diferentes métodos definem os limites e o teto das cavidades. O algoritmo implementado no KVFinder emprega uma definição de cavidade geométrica baseada na teoria da morfologia matemática (MATHERON, 1975; SERRA, 1982).

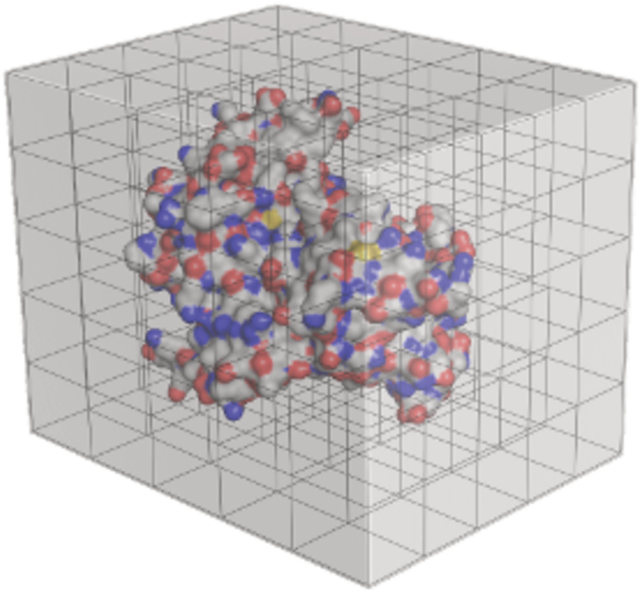


Figura 5: Representação esquemática da superfície molecular inserida na grade tridimensional. O KVFinder considera o raio de Van der Waals para os átomos que compõe a proteína a ser inserida na grade tridimensional.

A biomolécula de interesse é inserida em uma grade tridimensional subdividida em voxels. Cada voxel pode ser ocupado pela biomolécula analisada, considerando o raio de Van der Waals dos átomos, ou desocupado (Figura 5). Para garantir a acurácia e efetividade da detecção de cavidades, um sistema de dupla sonda, nomeadas de *Probe In* e *Probe Out*, foi implementado no KVFinder. O sistema de dupla sonda define duas superfícies moleculares com diferentes graus de acessibilidade, sendo os pontos de cavidade pertencentes à região onde não há sobreposição entre as superfícies das

duas sondas. Ao passo em que os pontos de cavidade são definidos, um algoritmo recursivo de busca em profundidade (*Depth-first search*; DFS) é aplicado para conectar os pontos pertencentes à mesma cavidade (TARJAN, 1972). Esse algoritmo considera cada ponto de cavidade como um nó, iniciando a busca por um nó acessível, todos os nós vizinhos separados por uma coordenada de grade são analisados. A busca continua até todos os pontos serem atribuídos a uma cavidade. Os pontos de cavidade são marcados como pertencentes a cavidade através de identificadores numéricos inteiros, os quais são atribuídos a pontos do mesmo procedimento de busca. (OLIVEIRA et al., 2014). O funcionamento desse sistema está ilustrado na Figura 6.

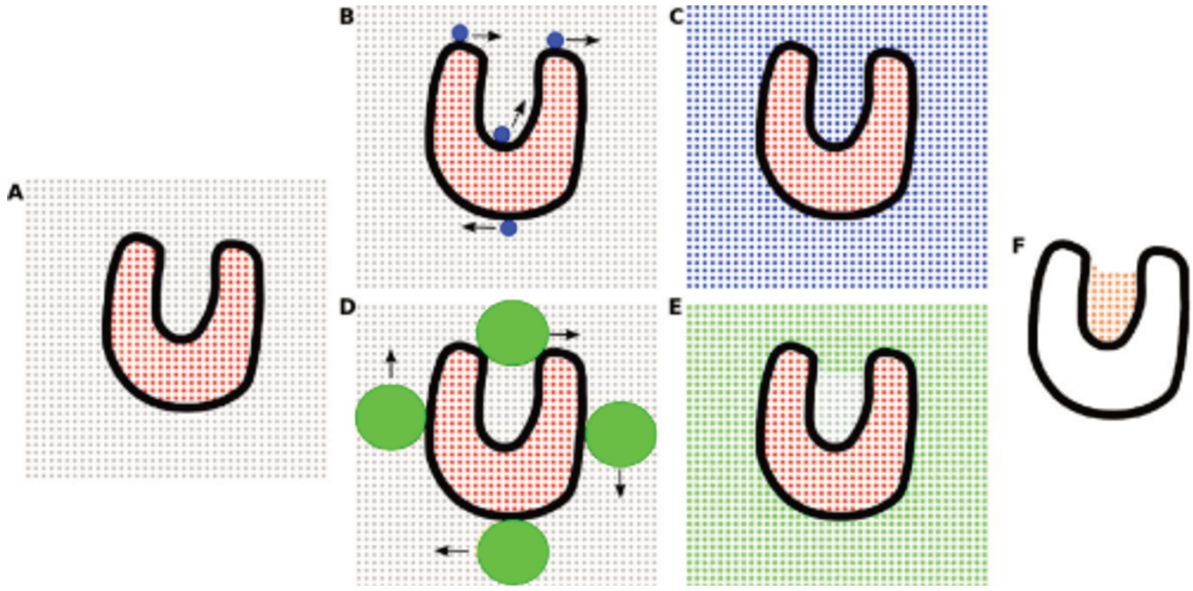


Figura 6: Representação didática do funcionamento do KVFinder. (A) Representação bidimensional da proteína em uma grade mostrando pontos da proteína (vermelho) e da cavidade (cinza). **(B)** Varredura da superfície pela sonda *Probe In* (azul). **(C)** Pontos de grade sobrepostos pela sonda *Probe In* são marcados em azul. **(D)** Varredura da superfície pela sonda *Probe Out* (verde). **(E)** Pontos da grade sobrepostos pela sonda *Probe Out* (verde). **(F)** A cavidade (laranja) é definida como região explorada pela sonda *Probe In* e não detectada pela *Probe Out.*

1.5.2 Metodologia para estimativa de volume e área superficial

Para cada cavidade biomolecular prospectada, o KVFinder realiza a caracterização espacial baseada nos voxels rotulados, a qual inclui a estimativa de volume e área superficial. O volume é estimado pela soma dos voxels pertencentes a cavidade multiplicada pelo volume do voxel, como apresentado na Equação 1. Para a estimativa da área superficial, os pontos de cavidades superficiais são determinados com base em um filtro simples. O filtro elimina os voxels pertencentes a cavidade que são cercados por pontos de cavidade, portanto, não é adjacente a nenhum ponto de

biomolécula. Os pontos não eliminados pelo filtro são os voxels superficiais. Por fim, a área superficial é estimada pela soma dos voxels de superfície multiplicada pela área de uma face, conforme apresentado na Equação 2.

$$\widehat{V_i} = N_{V,i} \cdot h^3 \quad (1)$$

$$\widehat{A_i} = N_{S,i} \cdot h^2 \quad (2)$$

Onde $\widehat{V_i}$ é o volume estimado da cavidade i, $N_{V,i}$ é o número de voxels pertencentes a cavidade i, $\widehat{A_i}$ é a área superficial estimada da cavidade i, $N_{S,i}$ é o número de voxels de superfície pertencentes a cavidade i e h é o espaçamento de grade.

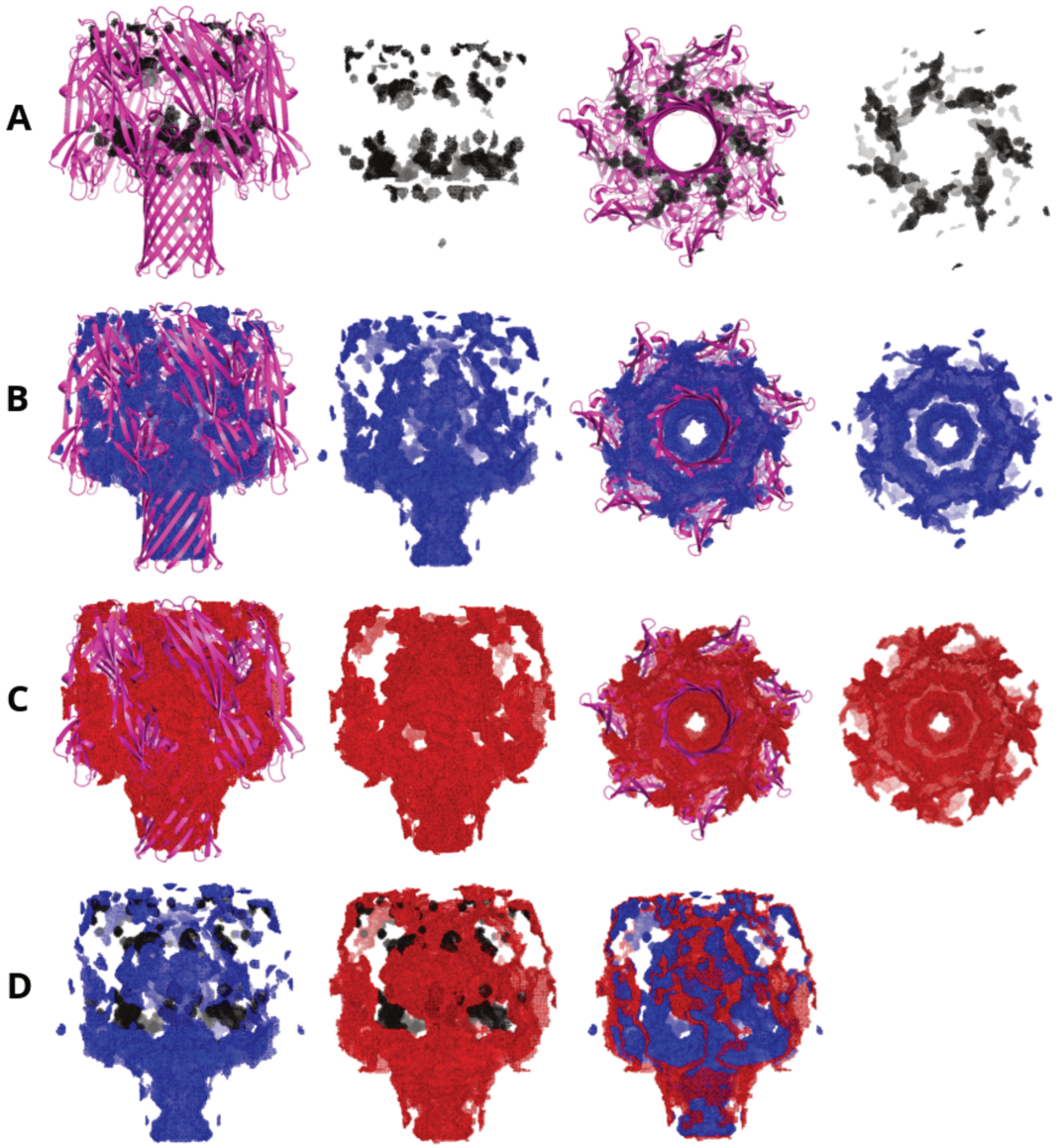
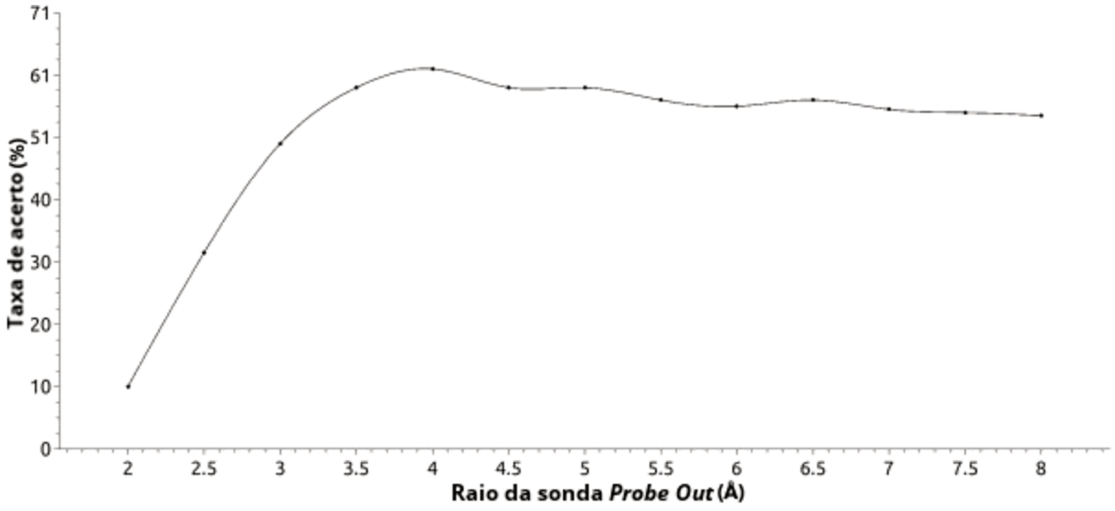


Figura 7: Prospecção de cavidades em uma toxina formadora de poros (PDB 4P24, rosa) para diferentes valores de sonda *Probe Out*. (A) 4,0 Å (preto). **(B)** 10,0 Å (azul). **(C)** 20,0 Å (vermelho). **(D)** Comparação pareada das cavidades prospectadas por diferentes valores de *Probe Out* através da sobreposição dos pontos de cavidade.

1.5.3 Desafios na detecção de cavidades superficiais

A detecção de cavidades biomoleculares superficiais é dependente do tamanho da sonda *Probe Out* escolhida. Prospecções com sondas *Probe Out* menores se limitam a encontrar cavidades internas. Com o aumento do tamanho da sonda, existe a prospecção de pontos mais superficiais nas cavidades e ao mesmo tempo, a identificação de cavidades superficiais (Figura 7). A topologia tem efeito determinante durante a prospecção de cavidades superficiais. A variação da topologia biomolecular dificulta a correta determinação dessas regiões, sendo que para cada região existe um valor mínimo de *Probe Out* a ser utilizado para sua prospecção. O aumento do tamanho da sonda *Probe Out* aumenta os recursos computacionais a serem utilizados pelo KVFinder e, por fim, impacta no tempo computacional de execução do programa. Além disso, os resultados presentes no artigo publicado por nosso grupo mostram como os valores de *Probe Out* alteram a taxa de sucesso da descoberta de sítios ativos (OLIVEIRA et al., 2014; Figura 8), sendo a taxa de sucesso determinada pela porcentagem de proteínas que tiveram o sítio do seu ligante encontrado pelo KVFinder pelo total de proteínas no banco de dados.



Fonte: Adaptado de Oliveira et al. (2014).

Figura 8: Taxa de sucesso da detecção de sítios ativos pelo KVFinder. O banco de dados de 198 alvos de drogas foi utilizado para determinação da taxa de sucesso em função dos valores de sonda *Probe Out*.

Tendo em vista a complexidade e escala do problema, o desenvolvimento e implementação de rotinas computacionais que utilizam computação paralela no KVFinder serão determinantes durante a prospecção de cavidades superficiais e cavidades com maior nível de detalhamento. As dificuldades geradas por diferenças

topológicas em biomoléculas podem ser contornadas utilizando grandes sondas *Probe Out*, pois o tempo de execução dessa tarefa em arquitetura paralela seria significativamente reduzido em comparação a execução serial.

1.5.4 Limitações computacionais

O sistema de dupla sonda usado no KVFinder implica na inserção da biomolécula estudada em duas grades tridimensionais distintas. O número de voxels de cada grade aumenta proporcionalmente com o nível de detalhamento da análise devido a redução do espaçamento da grade (Figura 4). No entanto, vale ressaltar que o tamanho médio da grade tridimensional é aproximadamente 150x150x150. A determinação das cavidades é baseada na análise de cada voxel e de sua vizinhança, sendo que o número de voxels a ser investigado na vizinhança depende do tamanho de sonda utilizada e do espaçamento da grade. Outro fator impactante na execução do programa são as dimensões da biomolécula estudada. Apesar de não ser uma variável controlável no programa, a mesma impacta no número total de voxels a serem utilizados na prospecção das cavidades, sendo que o espaçamento de grade precisa ser escolhido de forma a gerar um nível de detalhamento apropriado à análise. Por fim, o tempo de execução do KVFinder está diretamente relacionado com a escolha do espaçamento da grade tridimensional, com o diâmetro da sonda *Probe Out* e com as dimensões da biomolécula a ser inserida na grade.

A implementação atual do KVFinder é baseada em computação serial. A tarefa de detecção e caracterização espacial de cavidades biomoleculares é tratada como um problema computacional geométrico, gerando uma alta demanda de recursos computacionais. As operações geométricas realizadas pelo KVFinder são aplicadas em cada voxel da grade e os dados de cada voxel são independentes dos demais, possibilitando a implementação de computação paralela. A abordagem paralela consiste na distribuição das operações entre os múltiplos processadores ou núcleos de processamento ao invés de realizar cada operação sequencialmente. Essa redução no tempo de processamento permite a resolução de problemas biológicos de maior escala e complexidade.

1.6 Escala supramolecular

A escala supramolecular engloba moléculas, de maior escala e complexidade, resultantes da associação de espécies químicas, moléculas ou íons organizados em

entidades de alta complexidade, unidas por ligações metal-ligante e/ou ligações não-covalentes, como interações eletrostáticas, ligações de hidrogênio e forças de Van der Waals (NETTO; FREM; MAURO, 2008; WEBBER et al., 2016). A reversibilidade das interações de supramoléculas dão origem a complexos biomoleculares capazes de sentir e responder a estímulos fisiológicos ou imitar aspectos funcionais e estruturais da sinalização biológica (WEBBER et al., 2016).

Estudos biológicos e químicos recentes encontraram que supramoléculas são adequadas para o desenvolvimento de novas nanoestruturas e biomoléculas, mas também para o avanço em importantes campos da biomedicina, incluindo segmentação molecular, diagnóstico e tratamento de câncer, administração e liberação de fármacos, terapia de genes, regeneração de tecidos e regulação dinâmica de processos celulares (YOSHII et al., 2014; WEBBER et al., 2016; CAO; YANG; MAO, 2016). Por fim, o universo das supramoléculas é um campo vasto a ser explorado, devido ao vago conhecimento sobre as suas interações, e com uma alta gama de possíveis aplicações ainda a serem desenvolvidas.

1.7 Computação paralela

Existem duas formas básicas de computação: a computação serial e a computação paralela (Figura 9). O problema computacional é sempre dividido em partes discretas, sendo que cada parte é composta por um conjunto de instruções. A computação serial é a execução sequencial dessas partes, enquanto a computação paralela é o uso simultâneo de múltiplos elementos de processamento a fim de resolver o conjunto de instruções de cada parte do problema de forma concorrente (FOSTER, 1995). As principais vantagens da computação paralela são a redução de tempo, a resolução de problemas maiores e mais complexos, e a melhor utilização dos recursos do computador, sendo que os computadores atuais são arquiteturas paralelas com processadores multinúcleo (GRAMA et al., 2003). Por outro lado, os programas paralelos são mais difíceis de programar do que os sequenciais, pois a execução concorrente das tarefas produz novas fontes de erros, como a condição de corrida, que é uma falha na execução em que o resultado do programa é inesperadamente dependente da sequência ou sincronia das tarefas executadas (PATTERSON; HENNESSY, 1998). Além disso, a comunicação e sincronização entre as subtarefas paralelas também pode ser uma barreira para atingir um desempenho satisfatório em programas paralelos.

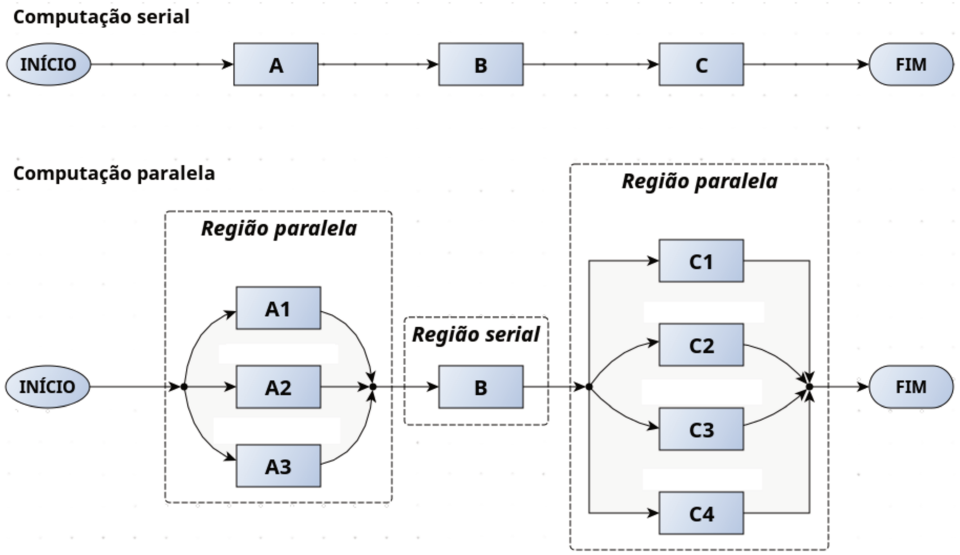


Figura 9: Representação esquemática da execução de um conjunto de instruções através de computação serial e paralela.

A compreensão prévia do problema computacional e do programa a ser abordado em paralelo é essencial para a paralelização. Primeiramente, os dados e a ordem de execução das instruções do problema devem ser independentes para a solução do problema. Por outro lado, o programa deve ser analisado a fim de localizar os *hotspots*, porções do código-fonte onde a maior parte do trabalho é realizado, e os gargalos, porções do código-fonte desproporcionalmente lentas (GRAMA et al., 2003). Por fim, tendo diagnosticado a independência dos dados, *hotspots* e gargalos do programa, é necessário aplicar o particionamento funcional e dos domínios. O particionamento funcional divide as instruções do problema em blocos ou ciclos concorrentes e dos domínios divide os dados em blocos independentes. Na abordagem paralela, ambos particionamentos podem ser aplicados para reduzir o tempo de execução (FOSTER, 1995; GRAMA et al., 2003). Entretanto, o particionamento ideal deve ser aliado ao balanceamento de tarefas, sendo que o balanceamento deve ocorrer de forma a distribuir quantidades aproximadamente iguais de trabalho entre as tarefas a fim de mantê-las ocupadas sempre (GRAMA et al., 2003).

O aumento da velocidade de um programa, conhecido como *speedup*, é uma medida relativa para analisar o desempenho entre dois sistemas processando o mesmo problema. O aumento ideal de velocidade é linear com o aumento de elementos de processamento. No entanto, poucos algoritmos são capazes de atingir

esse aumento ideal e a maioria deles possui um aumento praticamente linear para uma quantidade pequena de processadores, atingindo um comportamento assintótico para uma grande quantidade de elementos de processamento (Figura 10).

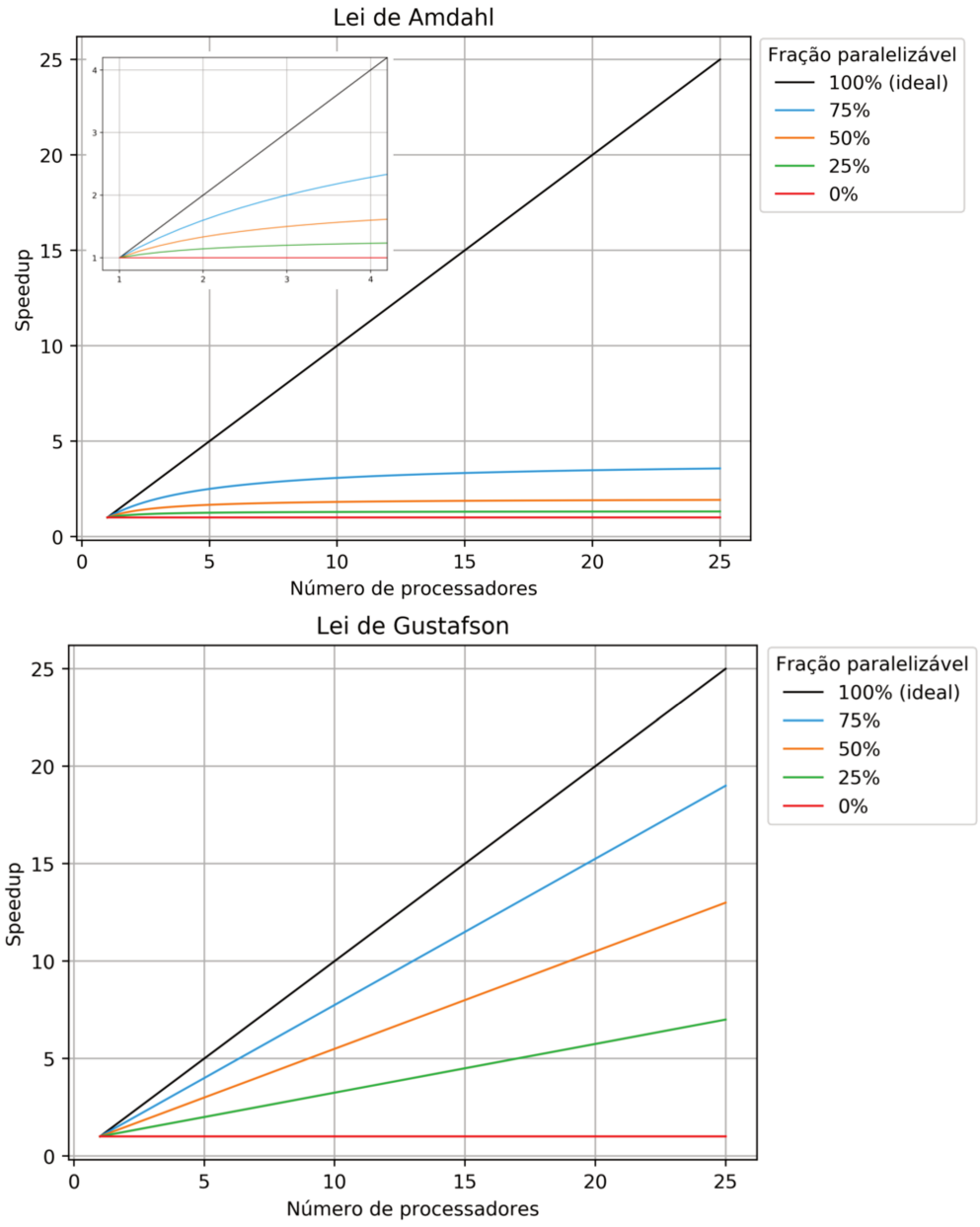


Figura 10: Representação gráfica do *speedup* de um programa. O *speedup* é apresentação em função do número de processadores disponíveis para diferentes frações de código paralelizável, considerando as estimativas das leis de Amdahl e Gustafson.

Esse aumento de velocidade pode ser descrito por duas funções: lei de Amdahl (AMDAHL, 1967) e lei de Gustafson (GUSTAFSON, 1988). A lei de Amdahl, Equação 3, assume o tamanho do problema computacional como fixo e a fração não paralelizável do programa como independente do número de processadores. A lei de

Gustafson, Equação 4, aborda as deficiências da lei de Amdahl, propondo que os programadores tendem a definir o tamanho dos problemas para explorar o poder computacional que se torna disponível à medida que os recursos melhoram. De certa forma, Gustafson (1988) redefine a eficiência, devido à possibilidade de que as limitações impostas pela fração sequencial de um programa possam ser combatidas pelo aumento de recursos computacionais disponíveis.

$$S_A = \frac{1}{f_{np} + \frac{f_p}{N_p}} \tag{3}$$

$$S_G = f_{np} + f_p \cdot N_p \tag{4}$$

Onde S_A é a estimativa de *speedup* pela lei de Amdahl, S_G é a estimativa de *speedup* pela lei de Gustafson, N_p é o número de processadores disponíveis, f_{np} é a fração não-paralelizável do programa e f_p é a fração paralelizável.

2 JUSTIFICATIVA

O programa KVFinder apresenta um conjunto de parâmetros customizáveis que possibilita a identificação de uma ampla gama de cavidades biomoleculares, desde bolsões profundos até fendas rasas na superfície. Porém, como qualquer método geométrico, exige muitos recursos computacionais e estruturas de dados que demandam memória (OLIVEIRA et al., 2014; DABERDAKU; FERRARI, 2016). Desta forma, a computação paralela de métodos geométricos baseados em grades é uma abordagem interessante para reduzir o tempo computacional, mantendo a acurácia. A implementação de rotinas paralelas no algoritmo do KVFinder viabilizará a análise de complexos supramoleculares em tempos menores, além de possibilitar a análise de dados em larga escala como, por exemplo, a prospecção e caracterização de sítios de ligação em várias proteínas de uma mesma família ou grupo, com o objetivo de encontrar a conservação e/ou compreender as funções bioquímicas desses sítios. A caracterização de cavidades por meio de descritores de propriedades físico-químicas, espaciais e constitucionais podem fornecer informações valiosas quanto as funções e interações das cavidades envolvidas em interações em qualquer escala molecular, incluindo a escala supramolecular. Desta forma, prospectar cavidades e caracteriza-las em aspectos físico-químicos, espaciais e constitucionais podem auxiliar no

desenho racional e aprimoramento de novos fármacos, além do melhor entendimento das funções bioquímicas e interações de complexos biomoleculares.

3 OBJETIVOS

3.1 Geral

Este trabalho tem como objetivo o desenvolvimento, otimização e implementação de uma nova versão do KVFinder, com rotinas de computação paralela, capaz de prospectar e descrever cavidades biomoleculares em estruturas de qualquer escala molecular.

3.2 Específicos

Os objetivos específicos se dividem em duas categorias: (1) otimização e atualização, e (2) desenvolvimento do conjunto de testes e de novas funcionalidades. A otimização e atualização do programa KVFinder incluem:

- Refatoração e otimização do código-fonte;
- Atualização do *plugin* gráfico e rotinas do KVFinder;
- Otimização da detecção de cavidades superficiais;
- Implementação de rotinas de computação paralela, parKVFinder (*parallel* KVFinder).

O desenvolvimento do conjunto de testes e de novas funcionalidades no algoritmo do KVFinder incluem:

- Desenvolvimento de um conjunto de estruturas proteicas representativas do PDB;
- Aprimoramento e implementação de descritores de propriedades nas cavidades prospectadas.

4 METODOLOGIA

O trabalho foi desenvolvido no Laboratório de Biologia Computacional (LBC), parte do Laboratório Nacional de Biociências (LNBio), no Centro Nacional de Pesquisa em Energia e Materiais (CNPEM), contando com a infraestrutura disponível no centro, incluindo um *cluster* de computação de alto desempenho com 15 nós e 92 *núcleos* de processamento. As etapas computacionais foram desenvolvidas, basicamente,

através das linguagens de programação C e Python, e do programa de visualização molecular PyMOL.

As etapas de desenvolvimento foram divididas em: atualização e otimização, implementação de computação paralela, melhorias incrementais, desenvolvimento de conjunto de testes e implementação de descritores de propriedades em cavidades biomoleculares.

4.1 Atualização e otimização

A versão do programa KVFinder publicada em Oliveira et al. (2014) foi desenvolvida para os sistemas operacionais Windows 7 e Ubuntu 12.04, e o programa PyMOL v1.6. As rotinas atualmente empregadas no programa KVFinder estão descritas na Figura 11. Como essas versões estão obsoletas, as atualizações do código-fonte e do *plugin* gráfico do KVFinder são necessárias para que a ferramenta tenha amplo acesso à comunidade científica. Além disso, também foi realizada a modernização da interface gráfica do *plugin* com o intuito de facilitar a utilização do programa por parte do usuário.

A implementação da nova versão do KVFinder, parKVFinder, utilizou as linguagens de programação ANSI C aliada à gama de bibliotecas *built-in* da linguagem. Os códigos foram compilados em ambiente UNIX, Windows e macOS, o programa apresenta compatibilidade com os sistemas operacionais Ubuntu 18.04, Windows 10 e macOS Mojave. A maior interatividade com o usuário final é garantida pelo KVFinder PyMOL *plugin*, desenvolvido em Python com *toolkit* gráfico Tk, integrando o parKVFinder com o programa de visualização PyMOL v1.8.

Anterior à implementação das rotinas de computação paralela, a compreensão e refatoração do código-fonte foram realizadas para aumentar a legibilidade por parte dos usuários e futuros desenvolvedores. Com a compreensão das rotinas do KVFinder, o escopo da otimização foi a remoção de redundâncias, a reestruturação de funções e estruturas de dados, a implementação de retentores (*buffers*) para a escrita de arquivos e a padronização dos arquivos de parâmetros (arquivos de configuração) e resultados. Além disso, a remoção de possíveis fontes de erro, como a variação das dimensões da grade tridimensional em função do espaçamento de grade e diâmetro da sonda *Probe Out* para a mesma biomolécula, também foi abordada.

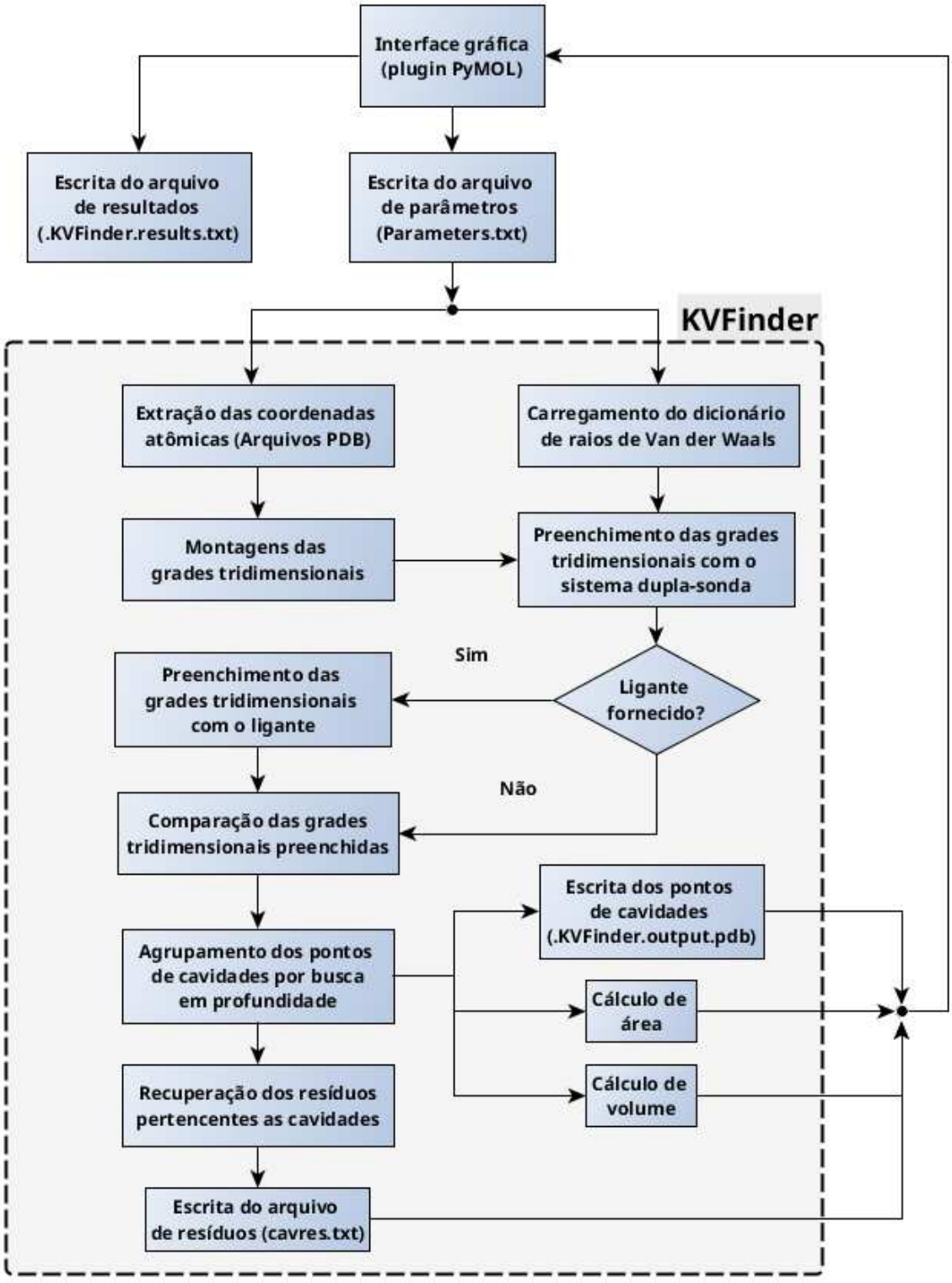


Figura 11: Fluxograma das rotinas empregadas no KVFinder. As rotinas atuam na prospecção e caracterização espacial de cavidades biomoleculares.

A refatoração é necessária para reduzir o tempo computacional de execução e identificar gargalos no algoritmo e a implementação de retentores de escrita reduz o

tempo computacional de escrita em disco. A reestruturação das estruturas de dados passa pela criação de registros (*structs* na linguagem C) capazes de acumular os parâmetros de entrada e os resultados finais. O registro dos parâmetros de entrada do parKVFinder possui um tamanho fixo na memória, não necessitando de uma estrutura de dados complexas, sendo apenas o agrupamento de inteiros, pontos flutuantes e cadeias de caracteres. Por outro lado, o registro dos resultados finais do parKVFinder possui tamanho variável por depender do número de cavidades encontradas pelo sistema de dupla sonda, exigindo a alocação dinâmica desse registro na memória conforme o número de cavidades encontrado. Além disso, os resultados como a lista de resíduos que cercam a cavidade encontrada necessitam de uma estrutura de dados especial, sendo agrupados na forma de lista encadeada com inserção ordenada de elementos.

A facilitação do uso e compreensão dos parâmetros usados pelo parKVFinder foram propiciados pela implementação de um arquivo de configurações padrão, como, por exemplo, TOML (*Tom's Obvious, Minimal Language*; https://github.com/toml-lang/toml), JSON (*JavaScript Object Notation*; https://www.json.org/) e YAML (*YAML Ain't Markup Language*; http://yaml.org/). Os formatos de serialização de código aberto, TOML, JSON e YAML, definem os parâmetros e as configurações iniciais de programas computacionais, sendo aplicáveis como arquivos de parâmetros e resultados do parKVFinder por apresentarem um sistema de chave-valor de fácil compreensão e edição. Estes formatos podem representar estruturas complexas de dados de uma forma simples, diferenciando variáveis, como números inteiros, números reais (pontos flutuantes), booleanas e cadeias de caracteres (*strings*), e estruturas, como vetores e tabelas. No entanto, a escrita de comentários no arquivo é possível apenas nos formatos TOML e YAML, que serve para documentar os dados passados aos arquivos, sem executar nenhum tipo de comando. Por fim, as diferenças entre os formatos de serialização consistem na diferença de sintaxe do arquivo e na escrita de comentários. Exemplos de arquivos de configuração nos formatos TOML, JSON e YAML estão apresentados no Anexo A.

As grades tridimensionais utilizadas no KVFinder são definidas por um conjunto de quatro pontos (P1, P2, P3 e P4) da biomolécula estudada e o espaçamento de grade. O ponto P1 representa os pontos mínimos dos eixos X, Y e Z, e os pontos P2, P3 e P4 representam os pontos máximos dos eixos X, Y e Z, respectivamente. As unidades da grade foram definidas pelas distâncias absolutas entre as posições

mínimas e máximas de cada eixo acrescido do tamanho da sonda nos dois sentidos do eixo e divididas pelo espaçamento de grade definido pelo usuário, conforme apresentado pelas equações 5, 6 e 7. Como as unidades dos eixos da grade só podem assumir valores inteiros, os valores de *m*, *n* e *o* são arredondados para o número inteiro superior. Portanto, as grades tridimensionais apresentam variação nas suas dimensões com a mudança dos parâmetros "espaçamento de grade" e "diâmetro da sonda *Probe Out*" para uma mesma biomolécula.

$$m = \frac{(x_2 - x_1) + d_p}{h} \tag{5}$$

$$n = \frac{(y_3 - y_1) + d_p}{h} \tag{6}$$

$$o = \frac{(z_4 - z_1) + d_p}{h} \tag{7}$$

Onde *m*, *n* e *o* são as unidades da grade nos eixos X, Y e Z, respectivamente, h é o espaçamento da grade, d_p é o diâmetro da sonda *Probe Out*, x_i, y_i e z_i são as coordenadas nos eixos X, Y e Z, respectivamente, no ponto Pi para i em {1, 2, 3, 4}.

No entanto, o espaçamento de grade não se mantém fixo e sofre pequenas oscilações devido ao modo que a grade e suas unidades são calculadas, gerando imprecisões numéricas associadas à detecção dos pontos de cavidades e caracterização espacial em uma mesma biomolécula. Essas imprecisões são vinculadas ao arredondamento das unidades de cada eixo de grade. Contudo, essas imprecisões numéricas podem ser mitigadas no parKVFinder pela utilização de unidades de grade múltiplas do espaçamento de grade, ou seja, o comprimento de cada eixo deve ser múltiplo do espaçamento de grade escolhido pelo usuário. Para que isso seja possível, a posição mínima de cada eixo será subtraída de um valor de forma que a posição resultante seja múltipla do espaçamento da grade e a posição máxima de cada eixo será adicionada de um valor de maneira que a posição resultante seja múltipla do espaçamento de grade. Por meio desta alteração, a grade tridimensional é padronizada para os parâmetros de entrada do programa parKVFinder, mitigando a sua principal fonte de imprecisão numérica.

4.2 Implementação de computação paralela

Primeiramente, a análise do perfil e desempenho computacional do programa KVFinder foi realizada com a ferramenta *Valgrind* (http://valgrind.org/), que é uma

estrutura de instrumentação para construir ferramentas de análise dinâmica. *Valgrind* possui um conjunto de ferramentas, cada qual executa algum tipo de tarefa de depuração, criação de perfil ou tarefas similares que ajudam a aprimorar os algoritmos. Para a análise de perfil e desempenho, a ferramenta de interesse é o *callgrind*, que é uma ferramenta de criação de perfil plano e gráfico de chamada. O perfil plano fornece o tempo de execução gasto em cada função e sua porcentagem em relação ao tempo total de execução. O gráfico de chamada apresenta, para cada função, qual função a chamou e qual foi chamada por ela. Desta forma, a ferramenta é capaz de determinar as funções gargalo e quais são as mais essenciais ao funcionamento do programa.

Em posse das funções gargalo, a independência dos dados executados em cada função foi avaliada para determinar a viabilidade da abordagem paralela em cada uma dessas funções. As funções que se mostraram paralelizáveis, foram paralelizadas por meio de uma interface de programação de aplicações (API) OpenMP (http://www.openmp.org/), baseada no modelo bifurcar-juntar para paralelizar a execução de programas em sistemas de multiprocessadores e sistemas de memória compartilhada para as linguagens de programação C, C++ e Fortran. A metodologia utilizada pelo API OpenMP consiste na implementação de múltiplos processos, sendo que um processo principal cria uma quantidade pré-definida de processos paralelos, conjuntos de dados ou instruções. Quando esse grupo de processos paralelos completa suas tarefas, eles são sincronizados e finalizados, mantendo apenas o principal. Os processos paralelos são executados concorrentemente, sendo alocados em diferentes processadores de acordo com o uso de recursos e balanceamento de tarefas na máquina. Além disso, o balanceamento das tarefas também foi aplicado nos processos paralelos criados pelo API para distribuir quantidades de trabalho iguais entre eles de forma a mantê-los ocupados a todo momento.

4.3 Melhorias incrementais no parKVFinder

Ainda que o parKVFinder tenha passado por técnicas de atualização, otimização e computação paralela, o programa ainda possui espaço para melhorias incrementais em suas rotinas. As melhorias incrementais são focadas no aumento da usabilidade e na redefinição de alguns conceitos do estado da arte do programa. As melhorias são: definição dos pontos de superfície, implementação do parâmetro

“distância de remoção”, implementação de método de definição indireta do espaçamento de grade e desenvolvimento de interface de linha de comando.

4.3.1 Definição dos pontos de superfície

O arquivo PDB de saída (<PDB>*.KVFinder.output.pdb*) contém os pontos internos de cavidades biomoleculares prospectadas pelo parKVFinder, porém o formato usado no KVFinder não diferencia os pontos de superfície dos demais pontos nas cavidades detectadas. Todos os pontos de cavidades são modelados como átomos de hidrogênio H (Figura 12A). A ausência da identificação dos pontos de superfície inviabiliza a visualização desses pontos no programa de visualização molecular PyMOL. Para superar essa limitação e melhorar a caracterização espacial da forma das cavidades, o formato do arquivo PDB de saída foi atualizado no programa parKVFinder para identificar os pontos de superfície. Os pontos internos permanecem sendo modelados como átomos de hidrogênio H e os pontos de superfície serão modelados como átomos HS (Figura 12B).

A

```
         1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
ATOM      1  H   KAA   259      -76.991 -20.140 -53.307   1.00  0.00
ATOM      2  H   KAA   259      -76.391 -20.140 -53.907   1.00  0.00
ATOM      3  H   KAA   259      -76.391 -20.140 -53.307   1.00  0.00
ATOM      4  H   KAB   259      -75.191 -29.140 -23.907   1.00  0.00
ATOM      5  H   KAB   259      -74.591 -29.740 -24.507   1.00  0.00
ATOM      6  H   KAB   259      -74.591 -29.740 -23.907   1.00  0.00
```

B

```
         1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
ATOM      1  H   KAA   259      -76.991 -20.140 -53.307   1.00  0.00
ATOM      2  HS  KAA   259      -76.391 -20.140 -53.907   1.00  0.00
ATOM      3  HS  KAA   259      -76.391 -20.140 -53.307   1.00  0.00
ATOM      4  H   KAB   259      -75.191 -29.140 -23.907   1.00  0.00
ATOM      5  HS  KAB   259      -74.591 -29.740 -24.507   1.00  0.00
ATOM      6  H   KAB   259      -74.591 -29.740 -23.907   1.00  0.00
```

Figura 12: Excerto de um arquivo PDB de cavidades biomoleculares. (A) Arquivo de saída atual com a representação de pontos internos. **(B)** Arquivo de saída novo com a representação de pontos de superfície e pontos internos. **Colunas 1-6:** nome do registro. **7-11:** número serial do átomo. **13-16:** nome do átomo. **18-20:** nome do resíduo (cavidade biomolecular, no caso). **23-26:** número identificador da sequência. **31-38:** coordenada ortogonal para X em ångströms. **39-46:** coordenada ortogonal para Y em ångströms. **47-54:** coordenada ortogonal para Z em ångströms. **55-60:** Ocupância. **61-66:** Fator de temperatura.

4.3.2 Implementação do parâmetro “distância de remoção”

A definição do teto de uma cavidade é uma dificuldade na detecção de cavidades biomoleculares. O KVFinder utiliza um parâmetro não-customizável, chamado “distância de remoção”, para a definição do teto das cavidades encontradas. Esse parâmetro elimina parte da região de fronteira com o meio das cavidades, sendo fixado a remoção de quatro unidades de grade em cada eixo. Desta maneira, o

parâmetro remove todos os pontos adjacentes a uma distância de quatro unidades de grade. A diminuição desse parâmetro possibilita a prospecção de cavidades mais superficiais; porém, ao mesmo tempo, dificulta a segregação das cavidades biomoleculares. Outro problema vinculado ao parâmetro é a distância de remoção ser dada em unidades de grade ao invés de um valor de comprimento em ångströms, sendo que a utilização de unidades de grade para o usuário é contra-intuitivo e a distância removida varia de acordo com o espaçamento de grade escolhido pelo usuário. Para solucionar essas limitações, a distância de remoção é um parâmetro customizável pelo usuário no parKVFinder e utiliza um valor de comprimento em ångströms em detrimento de valores de unidade de grade por ser mais intuitivo e simples. Assim, o valor em comprimento é dividido pelo espaçamento de grade para determinar o número de unidades de grade a ser removido na rotina. O funcionamento do parâmetro “distância de remoção” está esquematizado na Figura 13.

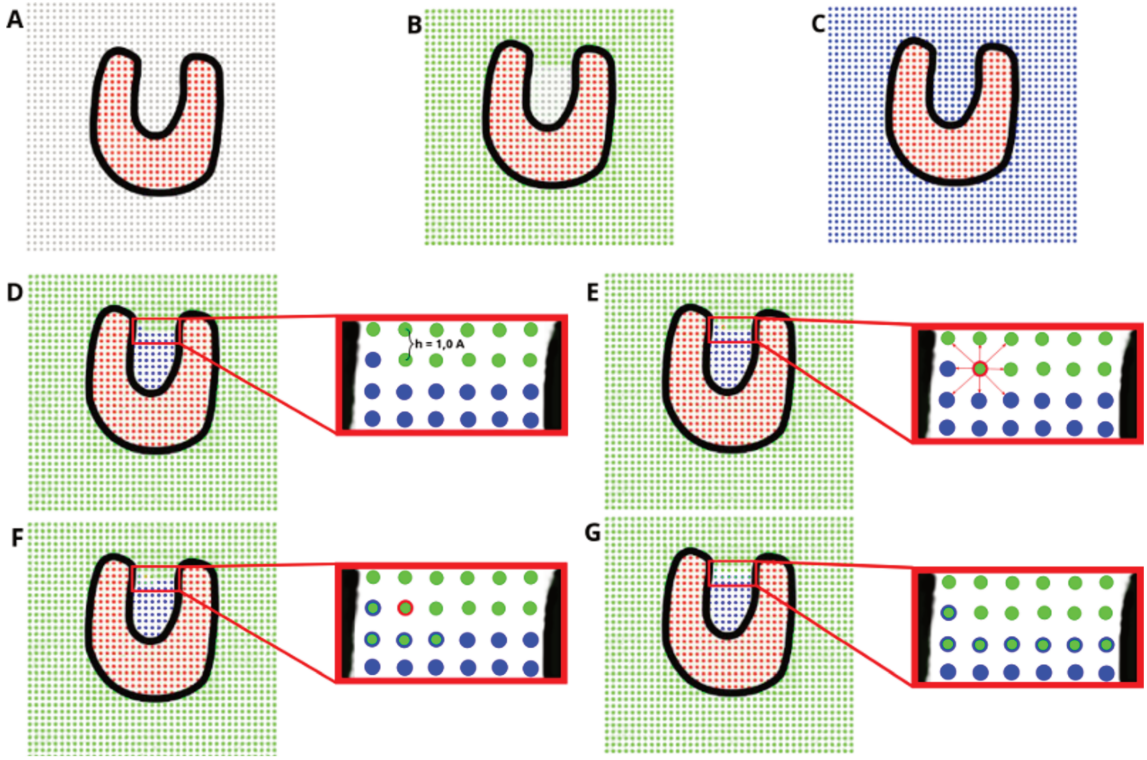


Figura 13: Representação esquemática do funcionamento do parâmetro “distância de remoção”. (A) Representação bidimensional da biomolécula em uma grade mostrando pontos da biomolécula (vermelho) e da cavidade (cinza). **(B)** Pontos da grade sobrepostos pela sonda *Probe Out* são marcados em verde. **(C)** Pontos de grade sobrepostos pela sonda *Probe In* são marcados em azul. **(D)** Os pontos de cavidades (azul) são definidos como diferença das superfícies das sondas e os pontos de meio (verde) são definidos como a intersecção das superfícies das sondas. O espaçamento de grade (h) e distância de remoção hipotéticos desse sistema são 1,0 ångström. **(E)** Partindo de um ponto de meio (verde) marcado em vermelho, todos os pontos adjacentes a ele são inspecionados. **(F)** Os pontos inspecionados que são pontos de cavidade (azul) são removidos da cavidade e passam a ser pontos de meio (pontos verdes marcados em azul). **(G)** O procedimento é repetido para todos os pontos da grade.

4.3.3 Implementação de método indireto do espaçamento de grade

A definição do espaçamento de grade pelo usuário não é intuitiva, ainda mais para usuários inexperientes em métodos baseados em grade. A implementação de uma alternativa à definição direta do espaçamento de grade facilita a utilização do programa parKVFinder por usuários iniciantes. A definição indireta é baseada no conceito de resolução de imagem, porém o método utiliza o volume dos voxels ao invés da área dos pixels para definir a resolução. O método indireto disponibiliza volumes pré-determinados de voxel (resoluções) para posteriormente determinar o espaçamento de grade.

4.3.4 Desenvolvimento de interface de linha de comando

A interatividade do programa parKVFinder com o usuário final depende da interface gráfica do programa de visualização molecular PyMOL. No entanto, esse tipo de interatividade limita análises em larga escala, sendo que não é possível interagir diretamente com o programa parKVFinder sem o arquivo de parâmetros. Para superar essa limitação, a interface de linha de comando integrada ao programa parKVFinder foi desenvolvida e aceita os mesmos parâmetros customizáveis disponíveis no arquivo de parâmetros e no *plugin* gráfico do PyMOL. Os parâmetros que não forem passados pelo usuário assumem os valores padrões definidos pelo programa, excetuando o caminho para o arquivo PDB alvo. Além disso, a interface de linha de comando também aceita o caminho para o arquivo de parâmetros.

A interface de linha de comando apresenta limitações na segregação do espaço por não ser possível visualizar as estruturas tridimensionais das biomoléculas. Para superar essa limitação, dois métodos de segregação do espaço a ser analisado foram desenvolvidos para a interface de linha de comando. O primeiro método, denominado caixa de busca customizado, utiliza um arquivo separado por *tab* com as coordenadas mínimas e máximas de cada eixo (Figura 14A) para construir a caixa de busca. Já o segundo método, denominado caixa de busca por resíduos, utiliza um arquivo separado por *tab* com uma lista de resíduos na forma número do resíduo na estrutura e caracteres identificadores da cadeia separados por um sublinhado (Figura 14B), e um valor espaçador (*padding*) em ångströms. O método recupera os valores mínimos e máximos do conjunto de resíduos passados, e o comprimento do espaçador é adicionado aos valores máximos de cada eixo e subtraído dos valores

mínimos em cada eixo. Em posse desses valores mínimos e máximos de cada eixo, a caixa de busca por resíduos é definida para segregação do espaço no programa.

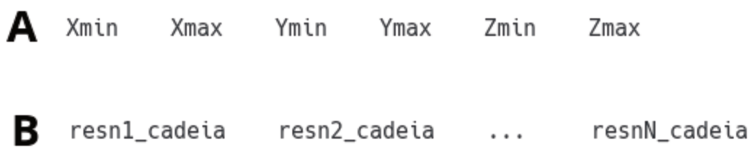


Figura 14: Arquivos padrões dos métodos de segregação do espaço na interface de linha de comando. (A) Caixa de busca customizada. **(B)** Caixa de busca por resíduos.

4.4 Desenvolvimento de conjunto de testes

O conjunto de testes foi desenvolvido para analisar o desempenho do programa parKVFinder, versão atualizada, otimizada e paralela do programa KVFinder, em termos de capacidade de detecção de cavidades e tempo de execução. O conjunto de testes de proteínas, denominado kv1000, é formado por um subconjunto do RCSB Protein Data Bank (http://www.rcsb.org), contemplando 1000 entidades proteicas que se assemelham nas características da população presente no RCSB PDB. Além disso, as entidades proteicas presentes no conjunto de testes apresentam uma distribuição semelhante à população nos quesitos massa molecular e classificação enzimática para inferir confiabilidade na capacidade do parKVFinder em prospectar e caracterizar cavidades proteicas na população proteica do RCSB PDB.

4.5 Implementação de descritores de propriedades

As cavidades prospectadas pelo programa parKVFinder serão descritas de acordo com as suas características espaciais, constitucionais e físico-químicas.

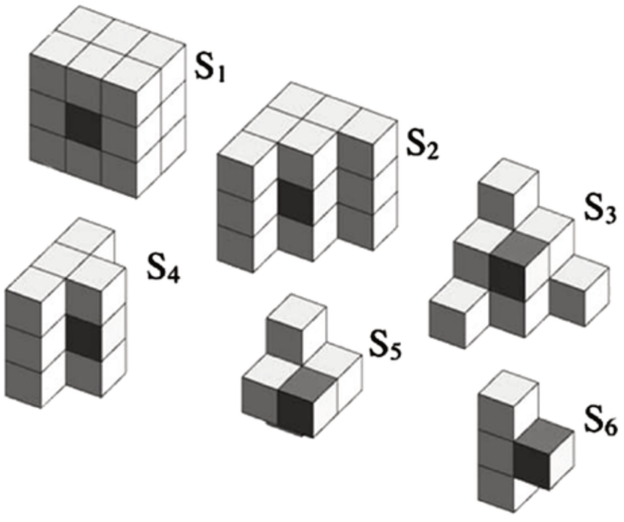
4.5.1 Descritores de propriedades espaciais

Os descritores de propriedades espaciais, volume e área, já estavam implementadas nas rotinas do programa KVFinder. O descritor de volume não foi alterado, porém foi avaliado seu desempenho no novo programa parKVFinder. Por outro lado, o descritor de área superficial foi atualizado com técnicas de estimativa de área superficial em objetos digitais compostos por voxels. Além de volume e área superficial, dois novos descritores espaciais foram agregados ao programa parKVFinder, o grau de profundidade, que pode ser calculado através da varredura dos pontos de cavidade em diferentes camadas em comparação ao meio, e o somatório da profundidade das cavidades encontradas.

4.5.1.1 Área superficial

O cálculo da área superficial das cavidades empregado no programa KVFinder considera apenas uma face do voxel como superfície, tornando a área superficial subestimada. Considerar apenas uma face como acessível aproxima a superfície real a uma combinação de superfícies retas, sendo o erro agravado para superfícies reais rugosas. Tendo em vista a redução deste erro entre a área real e a área estimada, um novo método para o cálculo da área superficial de cavidades foi implementado no novo programa parKVFinder.

A metodologia proposta é baseada na classificação dos voxels de superfície em seis classes propostas por Mullikin e Verbeek (1993) (Figura 15). Os voxels de superfície são divididos nas classes de acordo com a quantidade de faces em contato com a biomolécula e a localização destas faces. Os voxels do tipo $S_{1\text{-}3}$ aparecem em planos digitais e do tipo $S_{4\text{-}6}$ são encontrados em regiões curvas de fronteira. A área superficial (Equação 8) é estimada pela combinação linear da frequência das classes multiplicada pelo peso da classe. Tendo o esquema de classificação dos voxels definido, Mullikin e Verbeek (1993) definiram os pesos dessas seis classes, para remover o viés na estimativa da área para orientações aleatórias do plano e minimizar o erro quadrático médio. Os pesos são W_1 = 0,8940, W_2 = 1,3409, W_3 = 1,5879, W_4 = 2,0000, W_5 = 1,6667 e W_6 = 2,6667. De acordo com Windreich, Kiryati e Lohmann (2003), a metodologia proposta é de simples implementação, rápida computação e atinge uma boa acurácia. Além disso, a estimativa da área superficial tende a melhorar com o aumento da resolução, ou seja, a redução do tamanho dos voxels.



Fonte: Adaptado de Windreich, Kiryati e Lohmann (2003).

Figura 15: Classes de voxels de superfície.

$$\widehat{A_i} = \sum_{j=1}^{6} W_j \cdot N_{S,j,i} \quad (8)$$

Onde $\widehat{A_i}$ é a estimativa da área superficial da cavidade i, W_jé o peso da classe j de voxels de superfície e $N_{S,j,i}$ é o número de voxels de superfície da classe j na cavidade i.

4.5.1.2 Profundidade

A metodologia do descritor da profundidade implementado no novo programa parKVFinder utiliza o sistema de *flags* numéricas usada na grade tridimensional para identificação de cavidades biomoleculares. O sistema de dupla sonda determina os pontos de cavidade, pontos de biomolécula e pontos de meio com as *flags* 1, 0 e -1, respectivamente. Após o agrupamento dos pontos de cavidade pelo algoritmo DFS, os pontos de cavidades localizados a uma unidade de grade de um ponto de meio são os pontos de fronteira cavidade-meio, que são identificados como o valor negativo da *flag* numérica da cavidade correspondente. Para cada ponto da cavidade, as distâncias euclidianas entre o ponto e a fronteira cavidade-meio são calculadas. Por fim, a profundidade de cada ponto de cavidade é definida pela menor distância euclidiana entre o ponto e a fronteira. No entanto, existem cavidades que não possuem contato com o meio, ou seja, são bolsões enclausurados dentro da biomolécula (*voids*). Esse tipo de ponto tem sua profundidade definida como zero por não ser acessível por este método. Os valores de profundidade em cada ponto de cavidade serão projetados na estrutura de saída através do fator de temperatura. O arquivo de resultados do parKVFinder apresenta a profundidade máxima e a somatória das profundidades, definida como volume de profundidade, para cada cavidade encontrada.

4.5.2 Descritores de propriedades constitucionais

Os descritores de propriedades constitucionais do programa parKVFinder são a composição e características dos resíduos formadores das cavidades e a contagem desses resíduos.

4.5.2.1 Resíduos formadores das cavidades proteicas

A recuperação dos resíduos pertencentes às cavidades prospectadas proporciona a descrição desses resíduos em termos de localização e características. A rotina de recuperação dos resíduos empregada no KVFinder realiza a busca de pontos de cavidade dentro de um cubo de busca para o átomo do resíduo analisado. O cubo de busca é construído com base no diâmetro da sonda *Probe In* e o raio de van der Waals do átomo analisado. O cubo tem aresta equivalente ao dobro da soma do diâmetro da sonda e o raio do átomo, e o cubo é centrado nas coordenadas do átomo analisado. Vale ressaltar que todas as coordenadas cartesianas são transformadas em unidades de grade, e os valores mínimos e máximos do cubo são arredondados para o número inteiro inferior e superior, respectivamente. Desta maneira, quando um átomo encontra um ponto de cavidade dentro do cubo de busca, o resíduo deste átomo é identificado como formador da cavidade. Porém, os resíduos recuperados podem estar distantes da cavidade e não contribuírem para a formação da mesma. Para superar essa limitação, o espaço de busca passa a ser uma esfera de diâmetro equivalente à aresta do cubo (Figura 16), assim reduzindo o espaço de busca para recuperar resíduos mais próximos da cavidade.

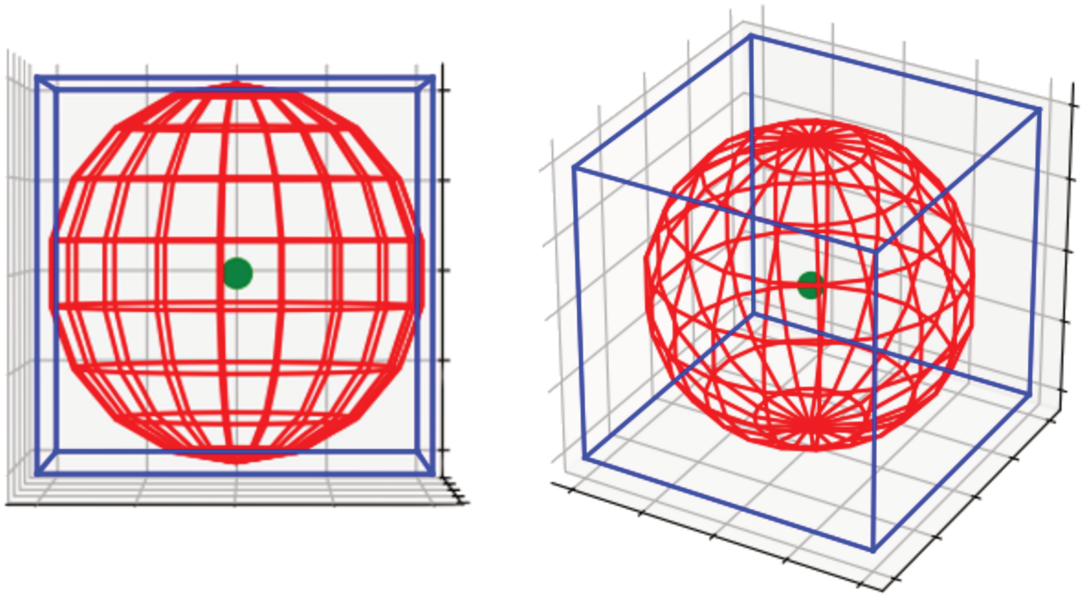


Figura 16: Representação dos espaços de busca de pontos de cavidades por átomos da estrutura biomolecular. O cubo de busca está representado em azul. A esfera de busca está representada em vermelho. O centro do espaço de busca, equivalente às coordenadas do átomo, está representado pelo ponto verde.

4.5.2.2 Composição, características e contagem dos resíduos formadores

Os vinte aminoácidos encontrados em proteínas diferem nas características estruturais e físico-químicas de suas cadeias laterais. De acordo com Lehninger, Nelson e Cox (1995), os resíduos podem ser agrupados em cinco classes, tendo como base as propriedades das suas cadeias laterais. As classes são: cadeias laterais (R1) não-polares e alifáticas, (R2) aromáticas, (R3) polares não-carregadas, (R4) carregadas negativamente e (R5) carregadas positivamente. No caso da presença de resíduos com modificações pós-traducionais, esses resíduos são agrupados em uma sexta classe de resíduos não-classificados. As classes e os aminoácidos presentes em cada classe estão apresentados na Figura 17.

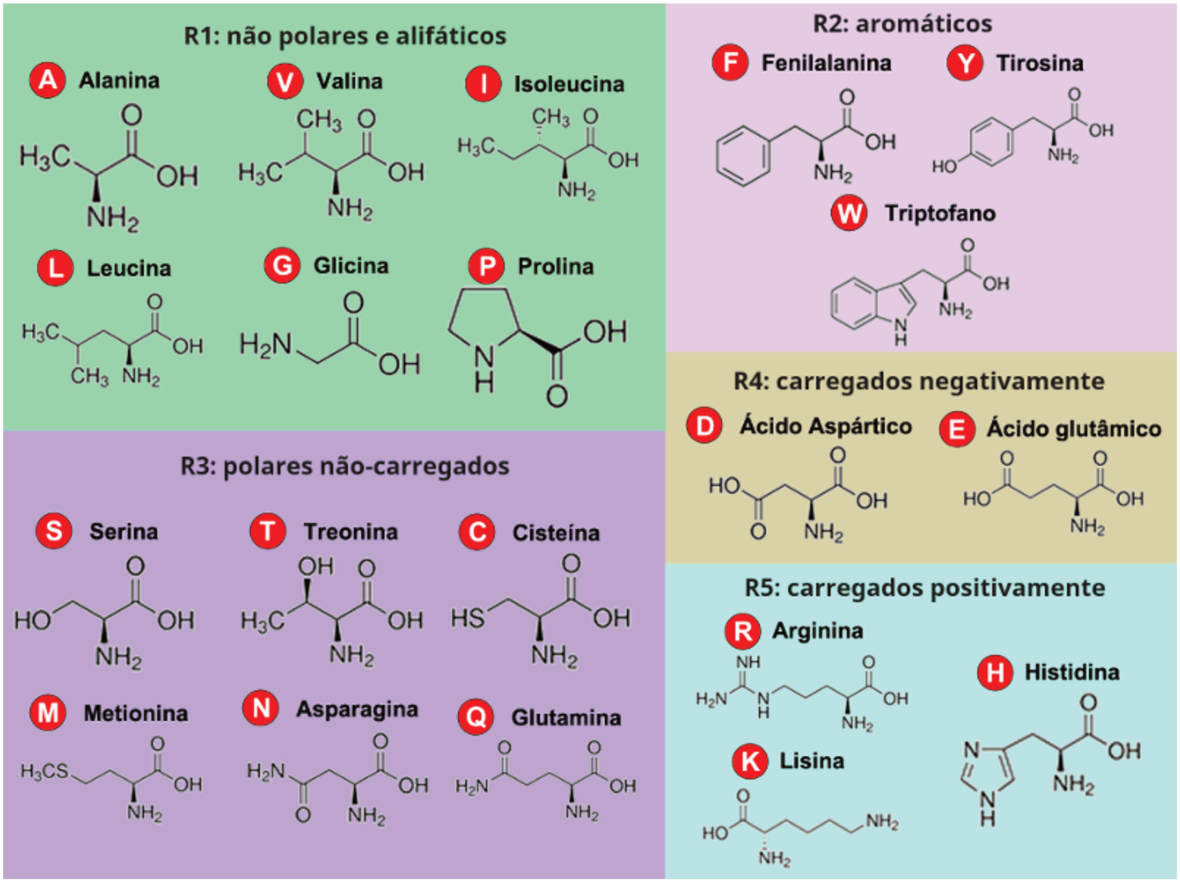


Figura 17: Classes de aminoácidos descrito em Lehninger, Nelson e Cox (1995).

Os resíduos formadores das cavidades são apresentados no arquivo de resultados do programa parKVFinder, apresentando a composição de resíduos e a contagem dos resíduos nas cavidades encontradas. Primeiramente, as informações de número do resíduo, caracteres identificadores da cadeia, nome do resíduo no código de uma letra e a classe do resíduo são apresentados para cada resíduo em

cada cavidade encontrada. Além disso, a quantidade de tipos e classes de resíduos presentes em uma cavidade são contabilizados, assim essas estatísticas dos resíduos são apresentadas por cavidade encontrada.

4.5.3 Descritores de propriedades físico-químicas

Ainda, novos descritores de propriedades físico-químicas também foram implementados no parKVFinder, dentre eles estão as escalas de hidrofobicidade e o potencial eletrostático.

4.5.3.1 Escalas de hidrofobicidade

O amplo estudo do conceito de hidrofobicidade foi responsável pela geração de diversas escalas. Os variados métodos para obtenção e parametrização são aplicados nas escalas, sendo eles: coeficiente de partição, técnicas de cromatografia, cálculos de área acessível ao solvente, mutação sítio dirigida e mensuração de propriedades físicas. No entanto, atualmente não existe um consenso a respeito de qual escala melhor representa o efeito hidrofóbico (BISWAS; DEVIDO; DORSEY, 2003), sendo assim, é dado ao usuário a opção de representar a cavidade obtida pelo parKVFinder de acordo com as seguinte escalas: Kyte & Doolittle (KYTE; DOOLITE, 1982), Wimley & White (WIMLEY; WHITE, 1996), Hessa & Heijne (HESSA et al., 2005), Zhao & London (ZHAO; LONDON, 2006) e Moon & Fleming (MOON; FLEMING, 2011) .

A metodologia das escalas de hidrofobicidade emprega um dicionário customizável de escalas de hidrofobicidade, nomeado *hidrophobicity_scales*, que contém o nome da escala, a referência para o artigo, o código de três letras do resíduo e seus respectivos valores na escala (Figura 18A). O dicionário tem suas informações carregadas no programa, sendo que cada escala carregada recebe um índice numérico (Figura 18B). Após o carregamento do dicionário pelo parKVFinder, uma grade tridimensional é criada para acumular os resíduos mais próximos dos pontos de superfície das cavidades. Os voxels da grade acumulam o código de uma letra do resíduo mais próximo. Para determinar o resíduo mais próximo dos pontos de superfície, as distâncias euclidianas entre os átomos do PDB e os pontos de superfície, considerando o raio de van der Waals do átomo e o diâmetro da sonda *Probe In*, são calculadas. E assim, o voxel do ponto de superfície acumula o código de uma letra do resíduo de menor distância para o mesmo. Por fim, os valores de cada

escala de hidrofobicidade fornecida pelo dicionário são recuperados no momento de exportar os arquivos PDB das cavidades de cada escala. Esses arquivos PDB (<PDB>.<*nome_da_escala*>*.output.pdb*) exportam apenas os pontos de superfície das cavidades com os valores de escala projetados através do fator de temperatura. O arquivo de resultados do parKVFinder apresenta o nome das escalas utilizadas na caracterização das cavidades como também os valores médios de cada escala para cada cavidade encontrada.

A

```
>HessaHeijne
# Recognition of transmembrane helices by the endoplasmic reticulum
translocon. Hessa T, Kim H, Bihlmaier K, Lundin C, Boekel J, Andersson
H, Nilsson I, White SH, von Heijne G. Nature. 2005 Jan 27;433(7024):
377-81, supplementary data.
ALA »   0.11
PHE »   -0.32
ILE »   -0.6
LEU »   -0.55
ASP »   3.49
TRP »   0.3
VAL »   -0.31
CYS »   -0.13
THR »   0.52
LYS »   2.71
GLN »   2.36
MET »   -0.1
TYR »   0.68
GLU »   2.68
SER »   0.84
ARG »   2.58
HIS »   2.06
ASN »   2.05
GLY »   0.74
PRO »   2.23
```

Nome da escala

Referência bibliográfica da escala

Valor da escala por resíduo

B

Índice	Escalas de hidrofobicidade
1	HessaHeijne
2	MoonFleming
3	KyteDoolitte
4	ZhaoLondon
5	WimleyWhite

Figura 18: Dicionário de escalas de hidrofobicidade. (A) Fragmento do dicionário nativo de escalas de hidrofobicidade com o valor da escala para cada resíduo. **(B)** Índices numéricos das escalas de hidrofobicidade no dicionário nativo.

4.5.3.2 Potencial eletrostático

Dentre os vários componentes da energética molecular, as propriedades de solvatação e as interações eletrostáticas são de grande importância devido ao alcance dessas interações. A aplicação de um modelo de solvatação contínuo é capaz de criar uma representação eletrostática qualitativa robusta, capaz de auxiliar químicos experimentais no desenho de fármacos. Para isso, o modelo de solvatação implícito baseado na PBE é aplicado (LAMM, 2003) para o cálculo do potencial eletrostático.

O descritor do potencial eletrostático do parKVFinder integra os pacotes PDB2PQR (DOLINSKY et al., 2004) e APBS (*Adaptative Poisson-Boltzmann Solver*; BAKER et al., 2001) para a resolução da PBE (JURRUS et al., 2017). Esses pacotes são distribuídos gratuitamente e com código livre, e amplamente utilizados pela

comunidade científica. O pacote PDB2PQR (DOLINSKY et al., 2004) estima os estados de protonação das biomoléculas, adiciona átomos de hidrogênio e alguns átomos pesados ausentes à estrutura cristalográfica, e assinala valores de carga e raio de acordo com o campo de força *Parse*, tendo o arquivo PDB como entrada e um arquivo PQR como saída do pacote.

O pacote APBS calcula o potencial eletrostático com cálculos de diferenças finitas configurado automaticamente pelo comando *mg-auto*. O arquivo PQR gerado pelo PDB2PQR serve como entrada do pacote APBS junto com um conjunto de dez parâmetros customizáveis. Os parâmetros incluem: (1) formulação da PBE a ser solucionada (PBE linear, não-linear ou forma linear regularizada), (2) condição de contorno para PBE (condição de contorno “zero”, Debye-Hückel simples ou Debye-Hückel múltiplo), (3) método de mapeamento das cargas pontuais da biomolécula na grade (discretização por splines lineares, beta-spline cúbico ou beta-spline quíntico), (4) modelos usados na construção das constantes dielétricas e acessibilidade de íons (*mol*, *smol*, *spl2* e *spl4*), (5) constante dielétrica do soluto, (6) constante dielétrica do solvente, (7) número de pontos de quadratura por ångström quadrado para usar em termos de superfície de cálculo, (8) raio das moléculas de solvente, (9) taxa de alteração para definições de superfícies baseadas em spline e (10) temperatura do sistema. Todas os parâmetros descritos acima possuem valores padrões, caso não sejam definidos pelo usuário.

A execução dos pacotes APBS-PDB2PQR é assíncrona à execução das rotinas do parKVFinder. A função *fork* cria um novo processo, processo filho, que é executado simultaneamente com o processo pai, programa parKVFinder. O processo filho executa as instruções dos pacotes APBS-PDB2PQR para o cálculo do potencial eletrostático. O processo pai, ao terminar de executar todas rotinas para detecção e caracterização de cavidades biomoleculares, excetuando o cálculo do potencial eletrostático, aguarda o processo filho terminar a execução de suas rotinas. A execução assíncrona só ocorre para o cálculo de potencial eletrostático, não ocorrendo em nenhuma outra seção do programa parKVFinder.

Após a execução dos pacotes APBS-PDB2PQR, o mapa eletrostático (<PDB>*.APBS.potential.dx*) é escrito no formato de dados escalares OpenDX. Esse arquivo é lido pelo programa e os dados de potencial são inseridos na grade tridimensional de forma direta, sem interpolação dos dados. Vale ressaltar que arquivos de potencial eletrostático em OpenDX podem ser passados diretamente ao

parKVFinder, não necessitando da execução dos pacotes APBS-PDB2PQR integrados ao parKVFinder, sendo inseridos na grade tridimensional da mesma forma. Por fim, os valores do potencial eletrostático são projetados nas cavidades encontradas no arquivo PDB de saída (<PDB>*.APBS.pdb*) através do fator de temperatura. O arquivo de resultados do parKVFinder apresenta o potencial eletrostático médio de cada cavidade encontrada.

5 RESULTADOS E DISCUSSÃO

5.1 Desenvolvimento do conjunto de testes

Para avaliação do desempenho computacional das novas versões do programa KVFinder, o conjunto de testes, kv1000, foi desenvolvido como um subconjunto representativo do repositório RCSB PDB. As estatísticas do repositório foram coletadas no dia 30 de janeiro de 2019, sendo esses dados utilizados como base para elaborar um subconjunto representativo do RCSB PDB. Nesta data, 148.243 estruturas estavam depositadas no repositório RCSB PDB, divididas em quatro tipos moleculares (Figura 19).

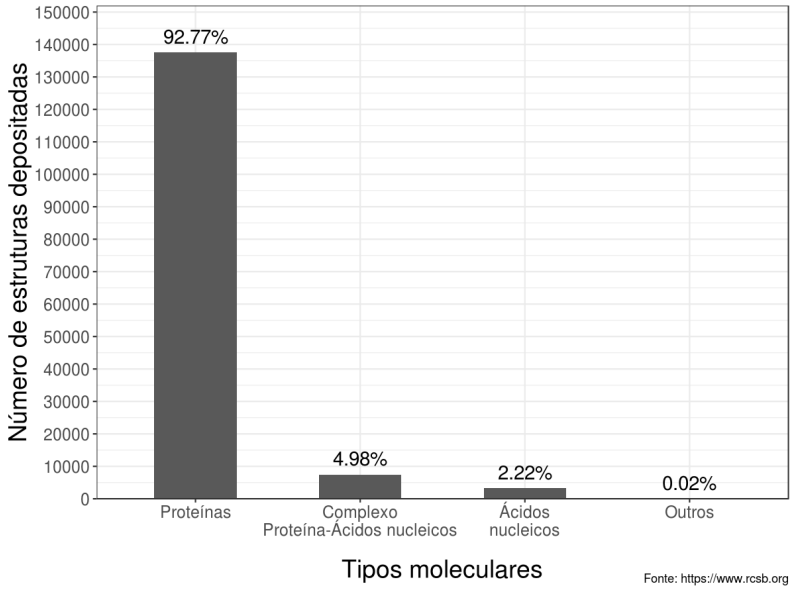


Figura 19: Distribuição de tipos moleculares no repositório RCSB PDB.

A partir do conjunto de 148.243 estruturas do RCSB PDB, são mantidas apenas as 137.525 (92,77%) estruturas correspondentes a proteínas. Destas estruturas proteicas, existem 70.909 (51,56%) estruturas classificadas como enzimas pelos

critérios do RCSB PDB e a distribuição dos tipos enzimáticos dentro desse subconjunto está apresentado na Figura 20.

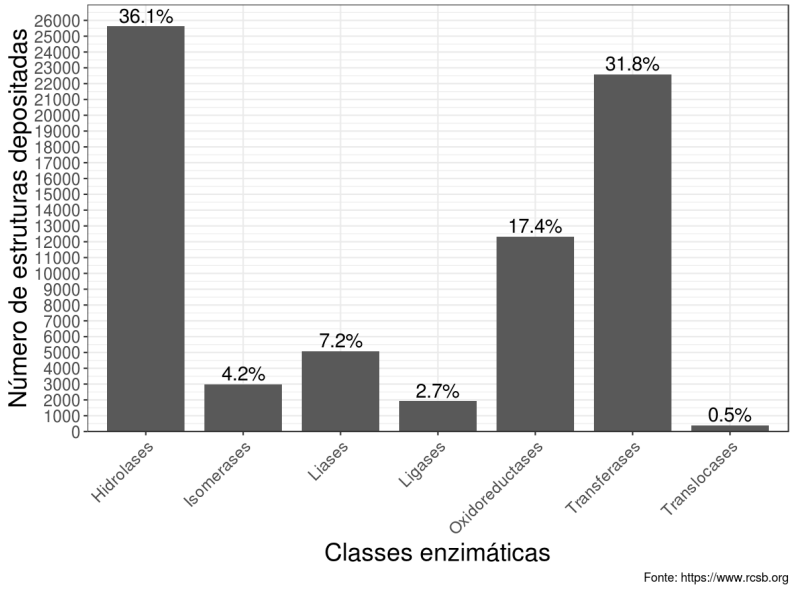


Figura 20: Distribuição de tipos enzimáticos no repositório RCSB PDB.

Destas estruturas contendo proteínas, o repositório determinou a existência de 252.809 entidades proteicas distintas. A distribuição de massa molecular das entidades proteicas distintas está apresentada na Figura 21. O banco de dados MUFOLD-DB (HE et al., 2014) realizou a filtragem das entidades distintas e recuperação de suas respectivas estruturas atômicas. A partir dessas entidades proteicas, o MUFOLD-DB manteve as entidades com similaridade menor que 30%, estruturas obtidas a partir da difração de raios X e massa molecular maior que 9 kDa. Esses critérios têm o intuito de eliminar estruturas proteicas com alta redundância em termos de similaridade sequencial e estrutural, baixa resolução estrutural e estruturas de possíveis ligantes. Desta maneira, obtém-se 12.315 (4,87% das entidades distintas) entidades proteicas com baixa similaridade e alta resolução.

As entidades distintas de baixa similaridade e alta resolução do MUFOLD-DB foram selecionadas aleatoriamente para compor o kv1000, seguindo um conjunto de regras como replicar as distribuições entre enzimas e não-enzimas (51,56% enzimas e 48,43% não-enzimas), de classes enzimáticas (Figura 20) e massa molecular das entidades (Figura 21B) do repositório RCSB PDB. Além disso, também foi definido um percentual máximo de entidades de um tipo de classificação não-enzimática. Como o maior subconjunto não-enzimático é a classificação “Sistema Imune”, e o mesmo

equivale a 3% do total de estruturas depositadas no repositório, os subconjuntos não-enzimáticos não foram maiores que 3% no conjunto de testes kv1000.

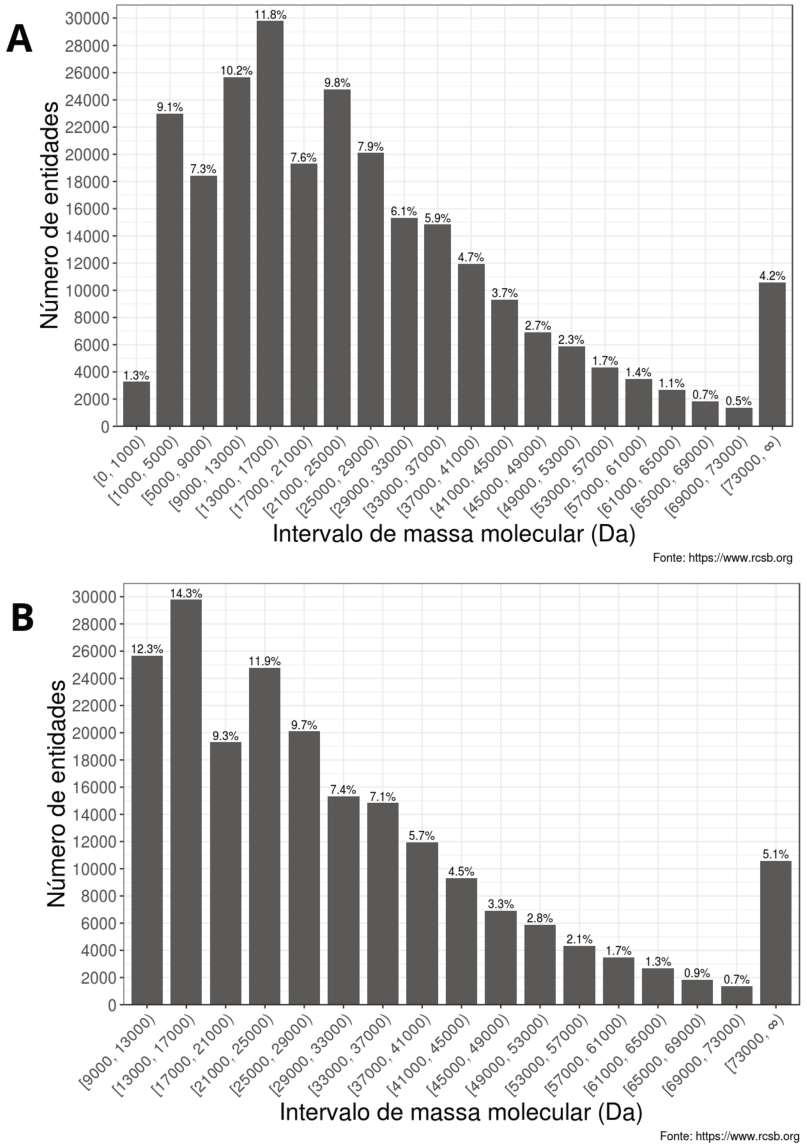


Figura 21: Distribuição de massa molecular das entidades distintas no repositório RCSB PDB. **(A)** Todas as entidades proteicas. **(B)** Entidades proteicas com massa molecular maior que 9 kDa.

Por fim, o conjunto de testes kv1000 é composto por 1000 entidades proteicas com baixa similaridade e alta resolução, sendo que as suas distribuições de massa molecular (Figura 22A), entre enzimas e não-enzimas (Figura 22B) e classes enzimáticas (Figura 22C) se assemelham as distribuições da população proteica do repositório RCSB PDB. Assumindo-se que as estatísticas das estruturas depositadas no repositório não variem significativamente no tempo, o conjunto kv1000 é considerado como representativo à população de estruturas proteicas depositadas no

RCSB PDB e os resultados obtidos para o conjunto podem ser expandidos para a população proteica do repositório RCSB PDB.

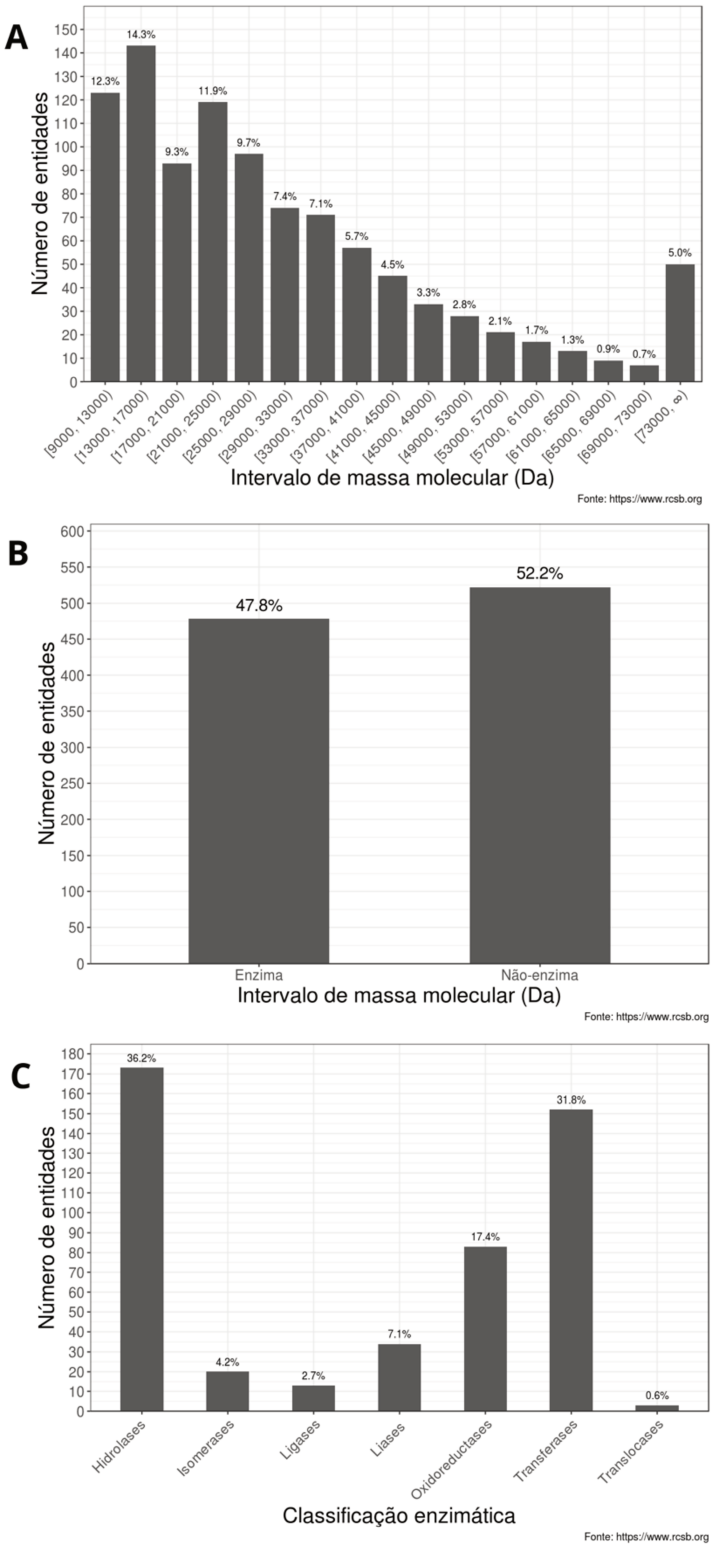


Figura 22: Estatísticas do conjunto de testes kv1000. (A) Distribuição de massa molecular das entidades. **(B)** Distribuição entre enzimas e não-enzimas. **(C)** Distribuição de classes enzimáticas.

5.2 Atualização, otimização e computação paralela das rotinas

O método baseado em grade tridimensional e esferas aplicado pelo KVFinder enquanto eficiente para detecção dos diferentes tipos de cavidades biomoleculares ainda possui espaço para otimização e melhoria das rotinas empregadas. O desempenho do programa KVFinder foi avaliado pela ferramenta *Valgrind* para determinar o perfil plano e o gráfico de chamada das funções. A ferramenta *Valgrind* requer a execução do programa para determinar o tempo consumido e quantidade de chamadas de cada função. Com base nisso, o programa KVFinder prospectou cavidades da grande subunidade do *Staphylococcus aureus* em complexo com lincomicina (PDB 5HKV, Figura 23) com os parâmetros padrões do programa exceto a sonda *Probe Out* com diâmetro de 8,0 ångströms. Os tempos computacionais, porcentagem do tempo total, quantidade de chamadas e descrição das principais funções do programa KVFinder estão na Tabela 2.

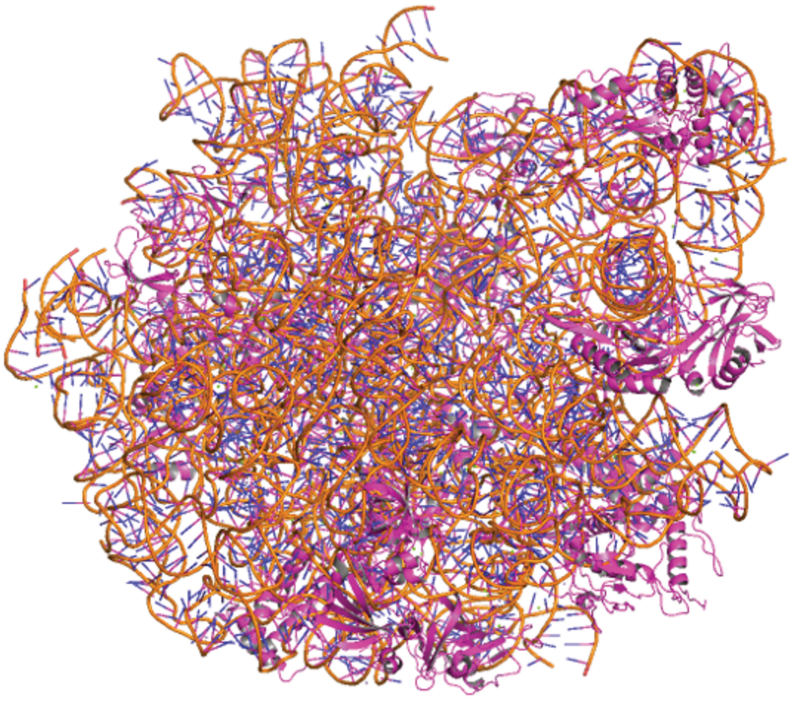


Figura 23: Estrutura da grande subunidade do *Staphylococcus aureus* em complexo com lincomicina. A estrutura do complexo (PDB 5HKV; 6304 aminoácidos) está representada na forma de *cartoon*.

De acordo com a Tabela 2, grande porção do tempo computacional de execução do KVFinder está concentrado em poucas funções, entre elas *Matrix_surf*, *Matrix_subtract* e *Matrix_fill2*. O foco da redução do tempo computacional passou pela

atualização, otimização e paralelização destas rotinas, mas sem esquecer das pequenas contribuições advindas das demais rotinas do KVFinder.

Tabela 2: Informações das funções do programa KVFinder pela ferramenta *Valgrind*. O tempo computacional e número de chamadas de cada função foram obtidos pela execução do programa para o PDB 5HKV.

Nome da função	Número de chamadas	Descrição	Tempo computacional (% do total)
Matrix_surf	2	Homogeiniza as superfícies moleculares geradas pelas sondas *Probe In* e *Probe Out*	7m57s (45,94%)
Matrix_subtract	1	Comparação das grades 3D preenchidas com sistema de dupla sonda	5m08s (29,71%)
Matrix_fill2	2	Preenchimento da grade 3D com a biomolécula estudada e a sonda escolhida (*Probe In* ou *Probe Out*)	1m44s (10,00%)
sqrt	6.641.978.101	Calcula a raiz quadrada positiva de um *double*	44s (4,28%)
check_pos	13.670.758	Verifica se os pontos vizinhos ao ponto analisado são pontos de biomolécula ou estão ao lado de um ponto que seja vizinho a um ponto de biomolécula	33s (3,15%)
DFS_search	1	Agrupa de pontos pertencentes a mesma cavidade por busca em profundidade	27s (2,56%)
remove_cavity	89	Remove a *tag* dos pontos das cavidades que não atingiram o volume de corte	25s (2,42%)
Matrix_export	1	Exporta os pontos de cavidades no formato de arquivo PDB	21s (2,02%)
Matrix_search	1	Recupera os resíduos em contato com as cavidades	4s (0,40%)
ceil	1.455.399.602	Retorna o número inteiro superior	4s (0,38%)
floor	748.583.069	Retorna o número inteiro inferior	2s (0,19%)
get_line	1.865.462	Extrai linha dos arquivos lidos	1s (0,12%)
Matrix_initialize	3	Atribui o valor inteiro 1 para todos os voxels da grade 3D	1s (0,07%)
Matrix_initialize2	2	Atribui o *double 0.0* para todos os voxels da grade 3D	1s (0,05%)
Matrix_surface	1	Aplica filtro para determinação dos pontos de superfície das cavidades	1s (0,05%)
Area_search	1	Realiza a estimativa de área superficial das cavidades	0,3s (0,03%)
main	**1**	**Programa KVFinder completo**	**17m18s (100%)**

De acordo com o fluxograma da Figura 11, o programa antigo apresenta algumas falhas lógicas nas rotinas empregadas. A preparação e escrita dos resultados do programa (<PDB>*.KVFinder.output.pdb* e <PDB>*.KVFinder.results.txt*) são contra-intuitivos pela escrita não ser realizada totalmente dentro do KVFinder e ser dependente do *plugin* gráfico. Logo, os arquivos de saída são distintos quando executamos o programa diretamente e via *plugin* gráfico do PyMOL, sendo que essa inconsistência prejudica a utilização do programa por parte do usuário final. Com base nisso, a preparação e escrita de arquivos passou a ser realizada totalmente pela nova versão do KVFinder, assim padronizando os arquivos de saída independente da forma de chamada do programa. Além disso, a etapa intermediária de escrita dos resíduos que cercam as cavidades no arquivo *cavres.txt* foi removida. Esse arquivo não é empregado em nenhuma rotina do programa, sendo utilizado apenas para escrever esses resíduos no arquivo de resultados, onde serão utilizados pelo *plugin* gráfico para identificar estes resíduos no PyMOL.

O arquivo de resultados (<PDB>*.KVFinder.results.txt*) e o arquivo de parâmetros (*Parameters.txt*) não são formatos de serialização que facilitam o uso e compreensão pelo usuário. Os modelos de escrita de arquivos de configuração padrão TOML, JSON e YAML foram testados e o formato TOML foi adotado como arquivo de resultados (<PDB>*.KVFinder.results.toml*) e arquivo de parâmetros (*parameters.toml*) por causa de sua simplicidade e facilidade para uso e compreensão por parte do usuário. Exemplos de arquivos de parâmetros e fragmentos dos arquivos de resultados da versão antiga (*Parameters.txt* e <PDB>*.KVFinder.results.txt*) e da nova versão do parKVFinder (*parameters.toml* e <PDB>*.KVFinder.results.toml*) estão apresentados no Anexo B.

A reestruturação das funções no procedimento de refatoração passou pela substituição de variáveis por ponteiros nas chamadas, reduzindo as etapas de duplicação de memória dentro do programa. Com a passagem de ponteiros, as variáveis globais podem ter seu valor alterado dentro das funções locais chamadas no programa em detrimento da atribuição do valor da variável local para a variável global. Além disso, as rotinas empregadas nas funções foram simplificadas para realizar menos atribuições e remover iterações desnecessárias, como a conversão de tipos de variáveis, leitura dos parâmetros de entrada, testes de decisão e redução do número de variáveis intermediárias. Desta forma, a redução e simplificação das atribuições realizadas, diminuíram o tempo computacional de execução do programa

Aliado a reestruturação de funções, a reestruturação das estruturas de dados para os parâmetros de entrada e os resultados finais simplificou a organização do código-fonte, o processamento de resultados parciais e o procedimento de escrita em disco. Por outro lado, a presença destes registros aumentou o consumo de memória, mas mantendo a eficiência e não prejudicando a execução do programa.

A refatoração também avaliou as rotinas empregadas quanto a redundância das tarefas executadas. Foram identificados testes de decisão para instruções que serão executadas independente das condições passadas ao programa. Sendo assim, estes testes de decisão podem ser removidos das rotinas do programa. Além disso, as funções *Matrix_initialize* e *Matrix_initialize2* também executam tarefas redundantes para o funcionamento do programa KVFinder, pois as grades tridimensionais já têm seu valor inicializado em cada voxel quando as mesmas são criadas. Além disso, a função *Matrix_initialize2* inicializa a grade tridimensional de voxels do tipo *double* e essa grade não é utilizada nas rotinas empregadas no programa. Porém, esta estrutura de dados permanece como legado da primeira versão do KVFinder (OLIVEIRA, 2011), a qual era utilizada para acumular valores de carga dos átomos da biomolécula estudada. Desta forma, ambas funções não são necessárias para o funcionamento do programa, sendo removidas na nova versão do programa.

O tempo de escrita em disco não aparece diretamente entre as principais funções do KVFinder (Tabela 2), mas as mesmas utilizam de funções para escrita em disco como, por exemplo, *fopen*, *fclose* e *fprintf*. Essas funções de escrita em disco são responsáveis por 7,5s (0,72%) da execução do KVFinder para a estrutura 5HKV. Para otimizar a escrita em disco, o registro dos resultados finais e os arquivos retentores de escrita foram implementados para a escrita dos arquivos de saída, sendo que a escrita é concentrada em uma única operação contínua ao final do programa devido as informações a serem escritas estarem salvas no registro criado. Os retentores atuam na escrita dos parâmetros essenciais para interpretação dos resultados, como caminhos para arquivos e “espaçamento de grade”, e dos descritores de propriedades para as cavidades encontradas (<PDB>*.KVFinder.results.toml*), as coordenadas dos pontos de cavidade (<PDB>*.KVFinder.output.pdb*) e a documentação da execução do programa KVFinder (*KVFinder.log*). Essas modificações são capazes de reduzir o tempo computacional desta tarefa para 1,2s (0,11%) da execução do KVFinder para a estrutura 5HKV.

5.2.1 Análise do desempenho computacional da atualização e otimização

Após os procedimentos para atualização e otimização do programa, as rotinas empregadas na nova versão atualizada e otimizada do programa KVFinder estão descritas na Figura 24.

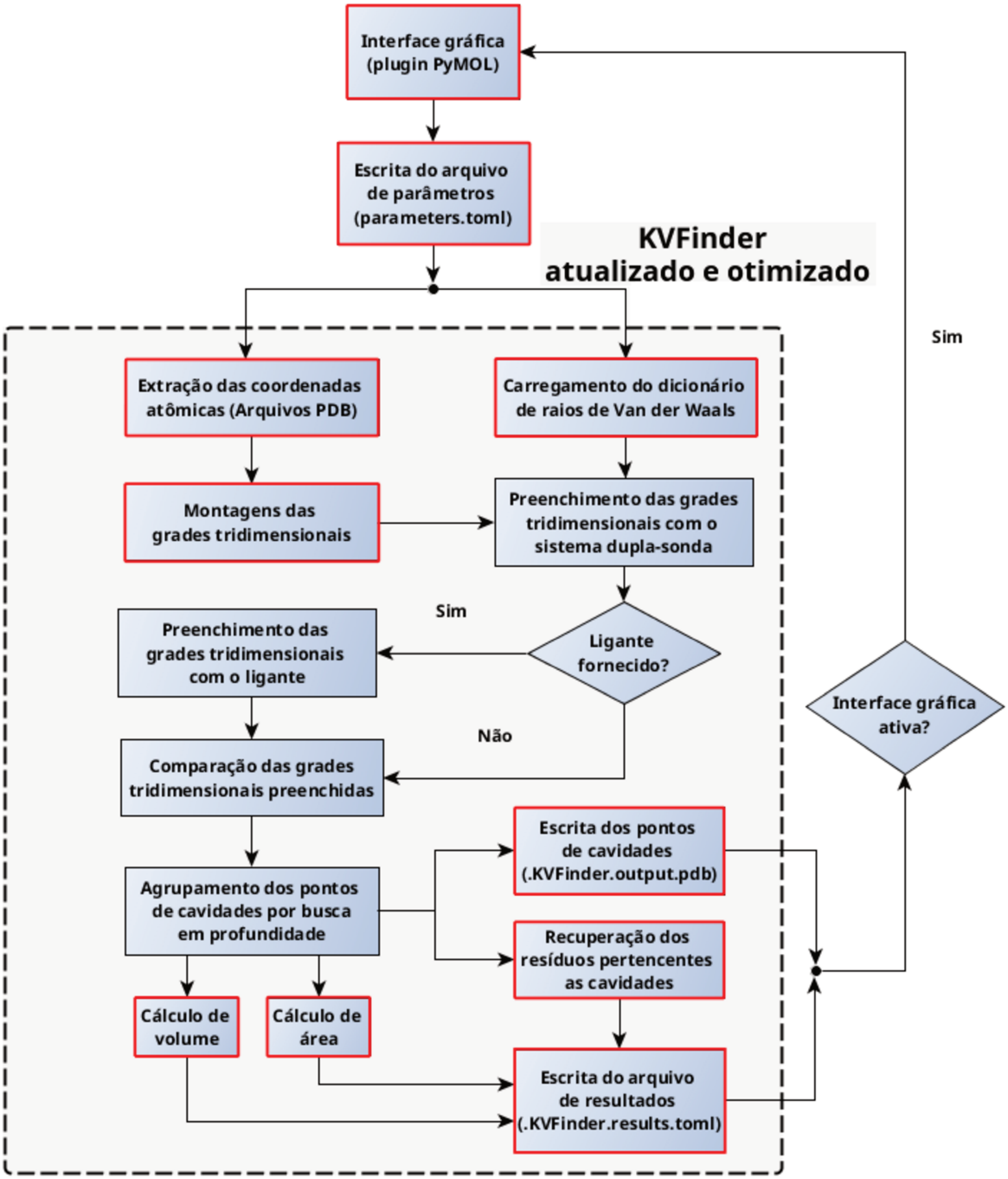


Figura 24: Fluxograma das rotinas empregadas no programa KVFinder atualizado e otimizado. As rotinas atuam na prospecção e caracterização espacial de cavidades biomoleculares. As rotinas alteradas estão identificadas em vermelho.

Para a análise de desempenho da atualização e otimização do programa KVFinder, é necessário definir o conceito de *speedup*. Partindo de uma mesma configuração computacional, o *speedup*, Equação 9, é a razão do tempo computacional entre dois programas que processam o mesmo problema. Neste caso, o *speedup* será a razão do tempo computacional do programa antigo pelo tempo computacional do programa atualizado e otimizado.

$$S = \frac{t_2}{t_1} \tag{9}$$

Onde S é o *speedup* entre os programas 1 e 2, t_1 é o tempo computacional de execução do programa 1 e t_2 é o tempo computacional do programa 2.

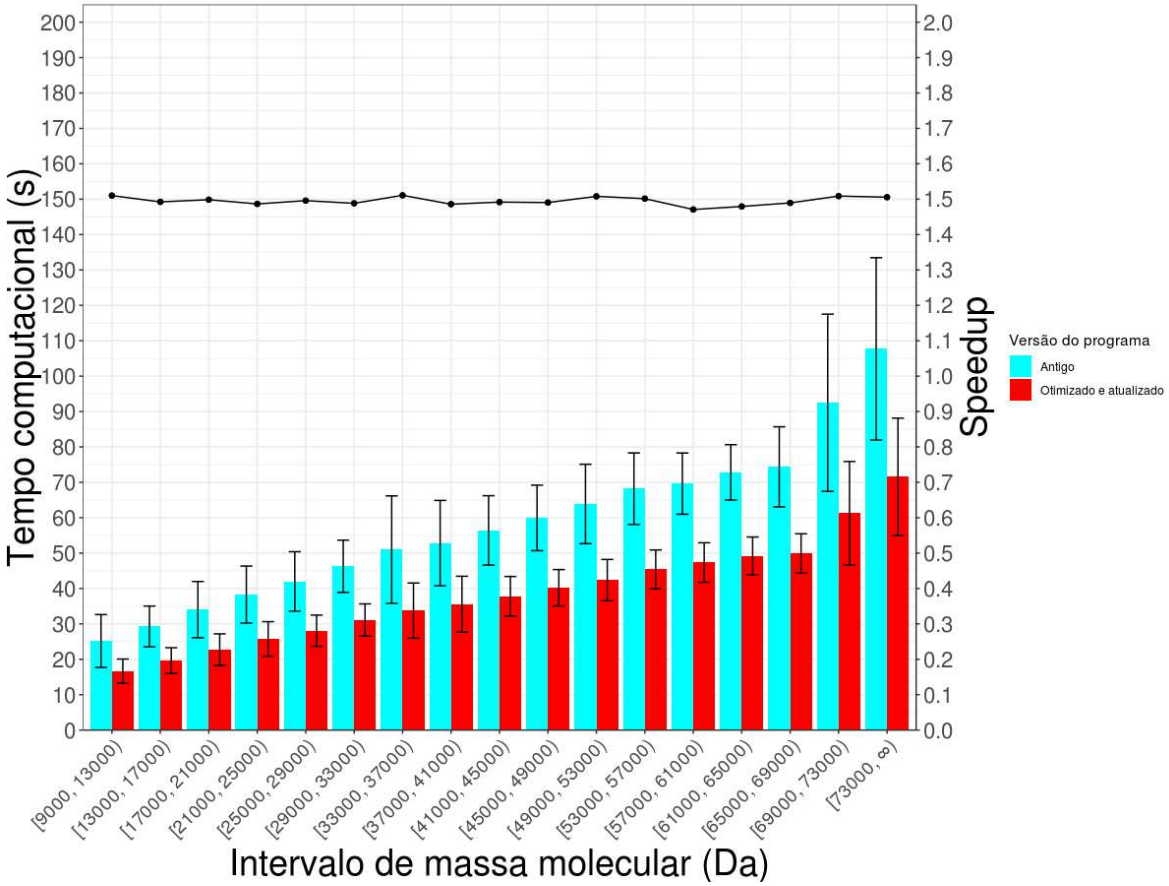


Figura 25: Desempenho computacional das versões do KVFinder. Gráfico de barras do tempo computacional de execução das versões do KVFinder otimizado e atualizado (vermelho) e antiga (ciano). O tempo computacional e o *speedup* estão apresentados para faixas de massa molecular das entidades proteicas do conjunto k1000. A escala do tempo computacional está apresentada no eixo da esquerda e do *speedup* no eixo da direita.

O novo programa KVFinder atualizado e otimizado e o programa KVFinder antigo foram executados em três replicatas para as 1000 entidades proteicas do conjunto de testes kv1000. Os tempos computacionais de cada programa e o

speedup, para cada intervalo de massa molecular, estão apresentados na Figura 25. Os tempos computacionais de execução da versão otimizada e atualizada do KVFinder foram reduzidos para todas as faixas de massa molecular do conjunto de testes, tendo um ganho de desempenho em torno de 1,5 vezes. Uma grande contribuição para o menor desempenho da versão antiga do KVFinder é a escrita dos arquivos temporários com os resíduos formadores das cavidades, sendo um arquivo temporário para cada cavidade encontrada.

Por fim, a versão otimizada e atualizada do KVFinder tem sua eficiência comprovada e a paralelização pode gerar um aumento significativo de desempenho para todas as faixas de massa molecular das entidades proteicas do conjunto kv1000. Portanto, a otimização e atualização do código-fonte para remover redundâncias, reestruturar funções e estrutura de dados, aumentar a legibilidade de usuários e simplificar os arquivos de entrada e saída do programa foi capaz de melhorar o desempenho global do KVFinder.

5.2.2 Análise do desempenho computacional da paralelização

O programa KVFinder possui operações matriciais em dados independentes, o que possibilita a paralelização dessas rotinas por meio da interface OpenMP. Conforme observado anteriormente na Tabela 2, o tempo computacional de execução do KVFinder está concentrado em poucas funções como *Matrix_surf*, *Matrix_subtract* e *Matrix_fill2*.

As funções *Matrix_surf* e *Matrix_subtract* possuem rotinas iterativas, na forma de estrutura de controle de repetição *for*, com dados independentes, sendo possível o acesso simultâneo dos voxels da grade e aplicação simultânea de funções nas informações contidas nesses voxels sem o comprometimento do resultado final do programa. A paralelização das estruturas de repetição é realizada por meio da diretiva de compilação *#pragma omp for collapse* do OpenMP. A alocação das instruções é feita de maneira dinâmica, ou seja, o balanceamento de tarefas será aplicado nos processos paralelos criados a fim de distribuir instruções entre elas de forma a mantê-las ocupadas a todo momento. Essa alocação dinâmica é realizada por meio da diretiva de compilação *#pragma omp for schedule(dynamic)* do OpenMP. Por outro lado, a função *Matrix_fill2* depende de uma estrutura de lista encadeada simples, impossibilitando a paralelização desta função, pois cada elemento da sequência é

armazenado em uma célula da lista e cada célula acumula o endereço da célula posterior.

Apesar das funções *Matrix_surf* e *Matrix_subtract* serem responsáveis por grande parte da carga computacional, as funções *remove_cavity* e *Matrix_filter* também possuem suas contribuições e podem ser paralelizadas. A função *remove_cavity* remove os identificadores numéricos das cavidades que não atingem o volume de corte escolhido pelo usuário. A função *Matrix_filter* seleciona os pontos a serem utilizados dentro do espaço de busca definido pelo usuário. Ambas funções realizam operações em dados independentes e as operações dentro delas independem da ordem que são realizadas. A paralelização das estruturas de repetição *for* também é pela diretiva de compilação *#pragma omp for collapse*. Sendo assim, o balanceamento de tarefas ocorre de forma a manter todos os processos ocupados e um conjunto de instruções não precisam esperar pela finalização de um conjunto anterior de instruções para iniciar. Então, o comando *nowait* é adicionado a diretiva anterior e a alocação dinâmica também é realizada pela diretiva de compilação *#pragma omp for schedule(dynamic)*. A definição das diretivas de compilação do OpenMP utilizadas no parKVFinder estão apresentadas no Anexo C.

A análise de desempenho da versão paralelizada do KVFinder depende dos conceitos de *speedup* e eficiência. O conceito de *speedup* já foi definido pela Equação 9 e, neste caso, será a razão do tempo computacional da versão serial pelo tempo computacional da versão paralela. O conceito de eficiência para *p* processos, Equação 10, é a porcentagem da razão do *speedup* para p processos pelo número de processos.

$$E_p(\%) = \frac{S_p}{p} \cdot 100\% \tag{10}$$

Onde E_p é a eficiência do programa para p processos, S_p é o *speedup* para p processos e p é o número de processos.

A nova versão atualizada, otimizada e paralelizada do programa KVFinder, parKVFinder, foi executada em três replicatas para as 1000 entidades proteicas do conjunto de testes kv1000, variando o número de processos desde 1 até 24 processos. Os tempos computacionais, o *speedup* e a eficiência para cada intervalo de massa molecular das entidades proteicas do conjunto de testes kv1000 no parKVFinder estão apresentados na Figura 26.

O programa parKVFinder apresentou uma significante redução no tempo computacional de execução para todas as faixas de massa molecular do conjunto de testes (Figura 26A). Esse aumento do desempenho computacional possibilita a solução de problemas maiores e mais complexos com o parKVFinder, pois o programa pode atuar na associação de entidades, desde pequenos complexos até biomoléculas supramoleculares, de diversas escalas, formas, estruturas e funções, e disponibiliza tempo computacional para o cálculo e aprimoramento dos descritores de propriedades espaciais, físico-químicas e constitucionais.

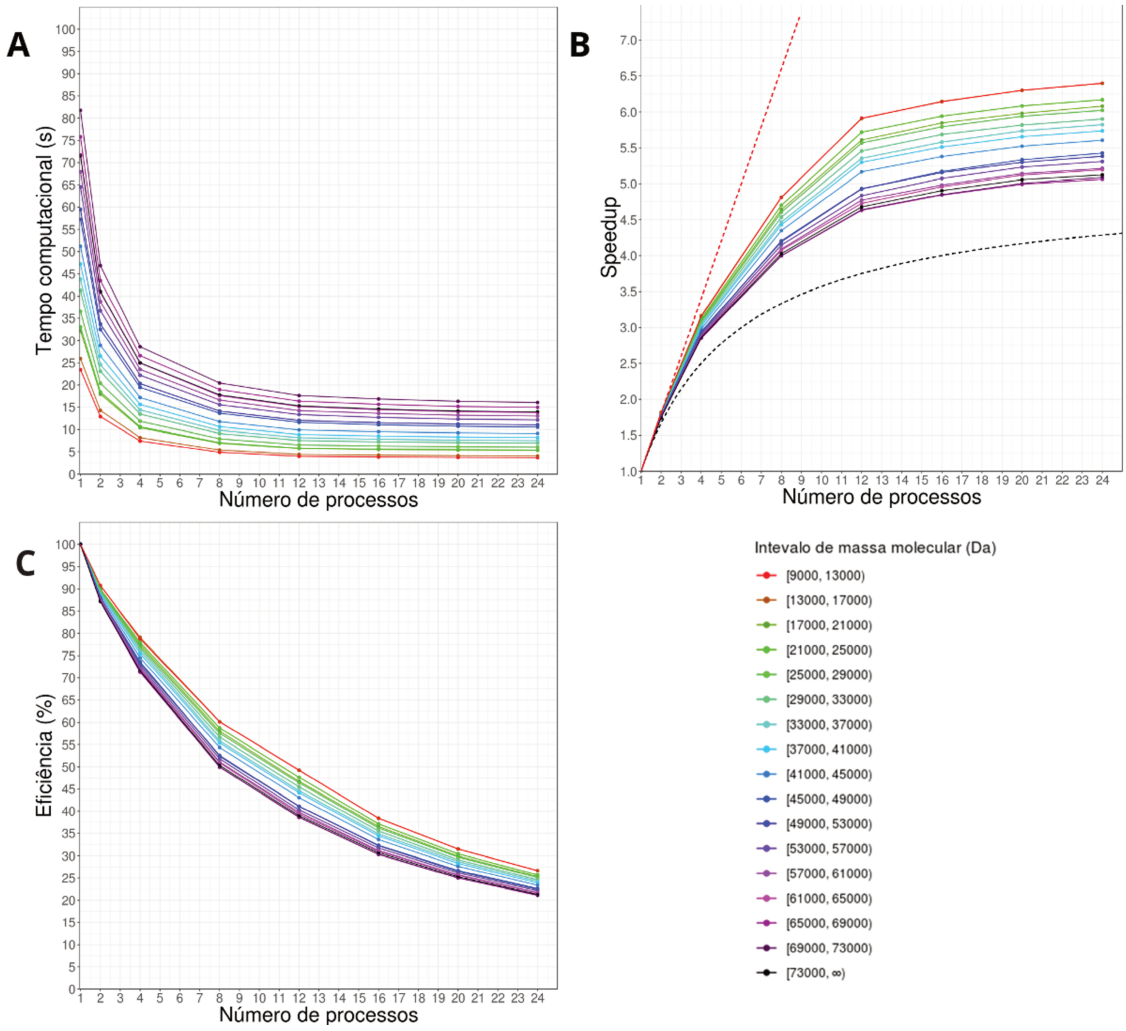


Figura 26: Desempenho computacional do programa parKVFinder. O programa foi executado em triplicata para cada entidade proteica do conjunto de testes kv1000. **(A)** Tempos computacionais em função do número de processos aplicados ao problema, por intervalo de massa molecular das entidades proteicas. **(B)** *Speedup* em função do número de processos aplicados ao problema, por intervalo de massa molecular das entidades proteicas do kv1000. A linha tracejada preta representa a Lei de Amdahl para um programa com uma fração paralelizável de 80% e a linha tracejada vermelha representa a Lei de Gustafson para um programa com um uma fração paralelizável de 80%. **(C)** Eficiência em função do número de processos aplicados ao problema, por intervalo de massa molecular das entidades proteicas do kv1000.

Para o caso de estudo da atualização, otimização e paralelização (Figura 23), aproximadamente 80% das funções executadas pelo programa são paralelizáveis (Tabela 2), porém esta fração é variável e dependente da biomolécula estudada. Com base na Figura 26B, a paralelização atingiu um comportamento praticamente linear para uma pequena quantidade de processos e um comportamento assintótico para uma grande quantidade de processos, conforme previsto pela Lei de Amdahl (Figura 10). O comportamento da paralelização no parKVFinder não obedece às considerações da Lei de Gustafson, pois o programa tem um problema de tamanho fixo para a mesma biomolécula e o aumento de recursos computacionais não acelera as partes sequenciais. Além disso, a eficiência da paralelização diminui de forma logarítmica com o número de processos utilizados Figura 26C, indicando que existe um limite para o ganho de desempenho com o aumento do número de processos aplicados ao programa. A contribuição de cada processo ao ganho de desempenho do programa diminui com o número de processos adicionados, indicando que as perdas por comunicação e distribuição de tarefas penalizam o ganho de processamento concorrente.

Portanto, a nova versão paralela do KVFinder, parKVFinder, foi implementada com rotinas de computação paralela, gerando um aumento de desempenho e melhor utilização dos recursos computacionais. Desta forma, possibilitando a inclusão e aprimoramento dos descritores de propriedades espaciais, físico-químicas e constitucionais no parKVFinder. Assim, melhorando a descrição e as possibilidades de análise de sítios de ligação em uma grande gama de estruturas biomoleculares.

5.3 Melhorias incrementais no parKVFinder

As melhorias incrementais implementadas no novo programa parKVFinder são: definição dos pontos de superfície, implementação do parâmetro “distância de remoção”, implementação de método de definição indireta do espaçamento de grade e desenvolvimento de interface de linha de comando.

5.3.1 Definição dos pontos de superfície

A definição dos pontos de superfície das cavidades é importante para a avaliação da forma e área superficial. O filtro de pontos adjacentes implementado no KVFinder considera os pontos de superfície como pontos de cavidade em que pelo menos um ponto adjacente é um ponto de biomolécula. O novo filtro de pontos

vizinhos implementado no parKVFinder considera os pontos de superfície como pontos de cavidade em que pelo menos um ponto vizinho é um ponto de biomolécula. Os conceitos de pontos adjacentes e vizinhos estão ilustrados na Figura 27.

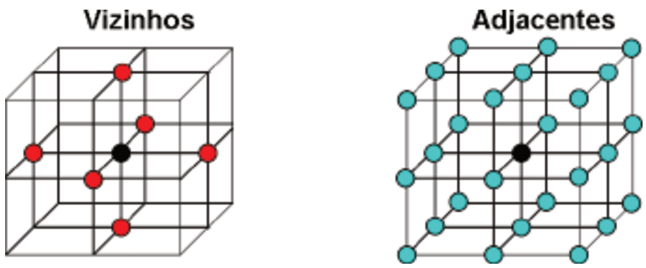


Figura 27: Representação esquemática de pontos vizinhos e pontos adjacentes.

Para visualizar as diferenças dos filtros de superfície, as cavidades da estrutura cristalina de uma subunidade da proteína quinase CAMP-dependente (PDB 1FMO) foram prospectadas usando os dois filtros: filtro de pontos adjacentes e filtro de pontos vizinhos. A comparação dos pontos de superfície do sítio de ligação da adenosina da proteína quinase estão apresentados na Figura 28.

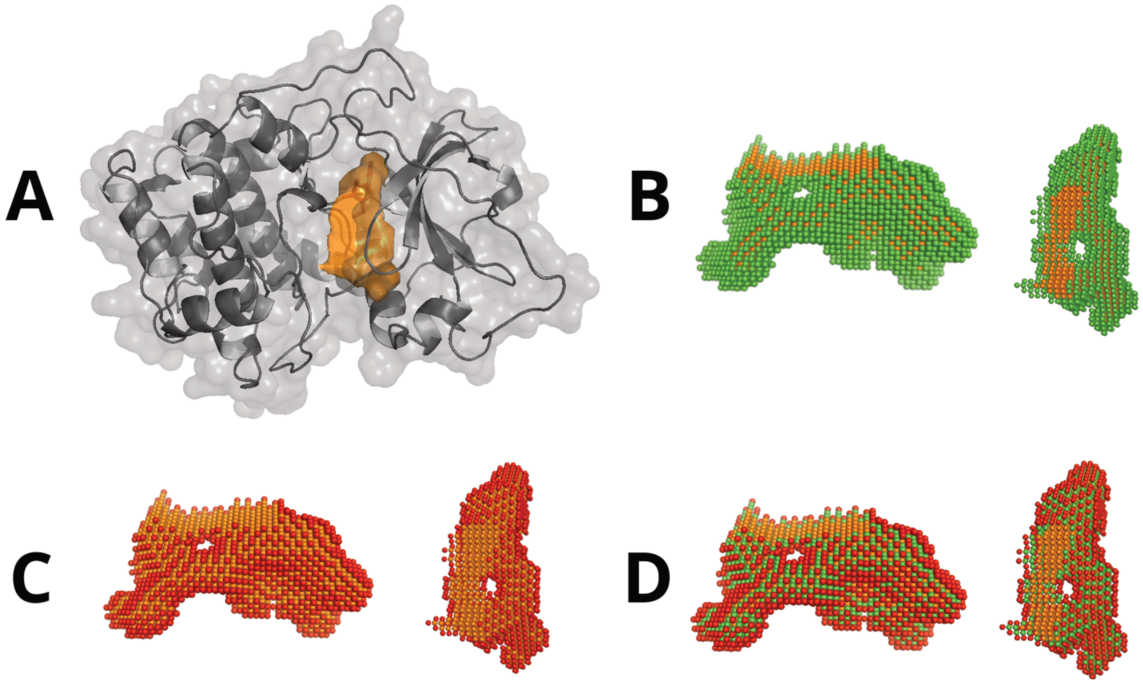


Figura 28: Comparação dos filtros de ponto de superfície. A comparação é feita para a cavidade do sítio de ligação da adenosina na proteína quinase A (PDB 1FMO). **(A)** Proteína 1FMO representada na forma de *cartoon* (cinza), cavidade do sítio de ligação da adenosina na forma de superfície (laranja) e adenosina na forma de *sticks* (vermelho). **(B)** Visão lateral (esquerda) e superior (direita) dos pontos de superfície (verde) e demais pontos da cavidade (laranja) do filtro de pontos adjacentes. **(C)** Visão lateral (esquerda) e superior (direita) dos pontos de superfície (vermelho) e demais pontos da cavidade (laranja) do filtro de pontos vizinhos. **(D)** Sobreposição dos pontos de superfície do filtro de pontos vizinhos (vermelho) aos pontos do filtro de pontos adjacente (verde) com os demais pontos de cavidades (laranja). Note que o filtro de pontos adjacentes (verde) mantém mais pontos de superfície do que o filtro de pontos vizinhos (vermelho).

A partir das comparações na Figura 28, é possível observar que o filtro de pontos vizinhos define menos pontos como pontos de superfície em comparação ao filtro de pontos adjacentes. Essa diferença na definição dos pontos de superfície indica que a área superficial das cavidades biomoleculares é superestimada no programa antigo como também a forma é imprecisa, pois os pontos de superfície do novo filtro são formados por uma camada com espessura de apenas um ponto de superfície, diferentemente do filtro aplicado no programa KVFinder. Além disso, como afirmado anteriormente (Figura 12B), os pontos de superfície serão marcados como átomos HS no arquivo PDB de saída e esses átomos também poderão ser visualizados como *nb_spheres* vermelhas no PyMOL para diferenciar os pontos de superfície dos demais pontos da cavidade. Portanto, o novo filtro de pontos vizinhos é mais eficiente para determinação dos pontos de superfície e a distinção dos pontos de superfície dos demais pontos de cavidade aprimora a visualização da forma das cavidades usando átomos HS para pontos de superfície e átomos H para os demais pontos da cavidade.

5.3.2 Implementação do parâmetro "distância de remoção"

O parâmetro não-customizável de distância de remoção era definido por quatro unidades de grade em cada direção independente do espaçamento de grade escolhido pelo usuário. No novo programa parKVFinder, o parâmetro é customizável pelo usuário e recebe um valor de comprimento em ångströms ao invés de unidades de grade, sendo menos dependente do espaçamento de grade escolhido.

O parâmetro "distância de remoção" ajuda a definir o teto das cavidades para um mesmo valor de espaçamento de grade e diâmetro da sonda *Probe Out*. Para avaliar a influência da distância de remoção na definição do teto da cavidade e prospecção de cavidades superficiais, as cavidades da estrutura 1FMO foram prospectadas para diferentes valores de distância de remoção, mantendo os pontos de cavidade que estão a menos de 8,0 ångströms do ligante, adenosina ou inibidor PKI, e usando o diâmetro da sonda *Probe Out* de 8,0 ångströms e espaçamento de grade de 0,6 ångström. A comparação do tetos da cavidade do sítio de ligação da adenosina e da capacidade de prospecção da cavidade superficial do sítio de ligação da PKI na estrutura 1FMO estão na Figura 29.

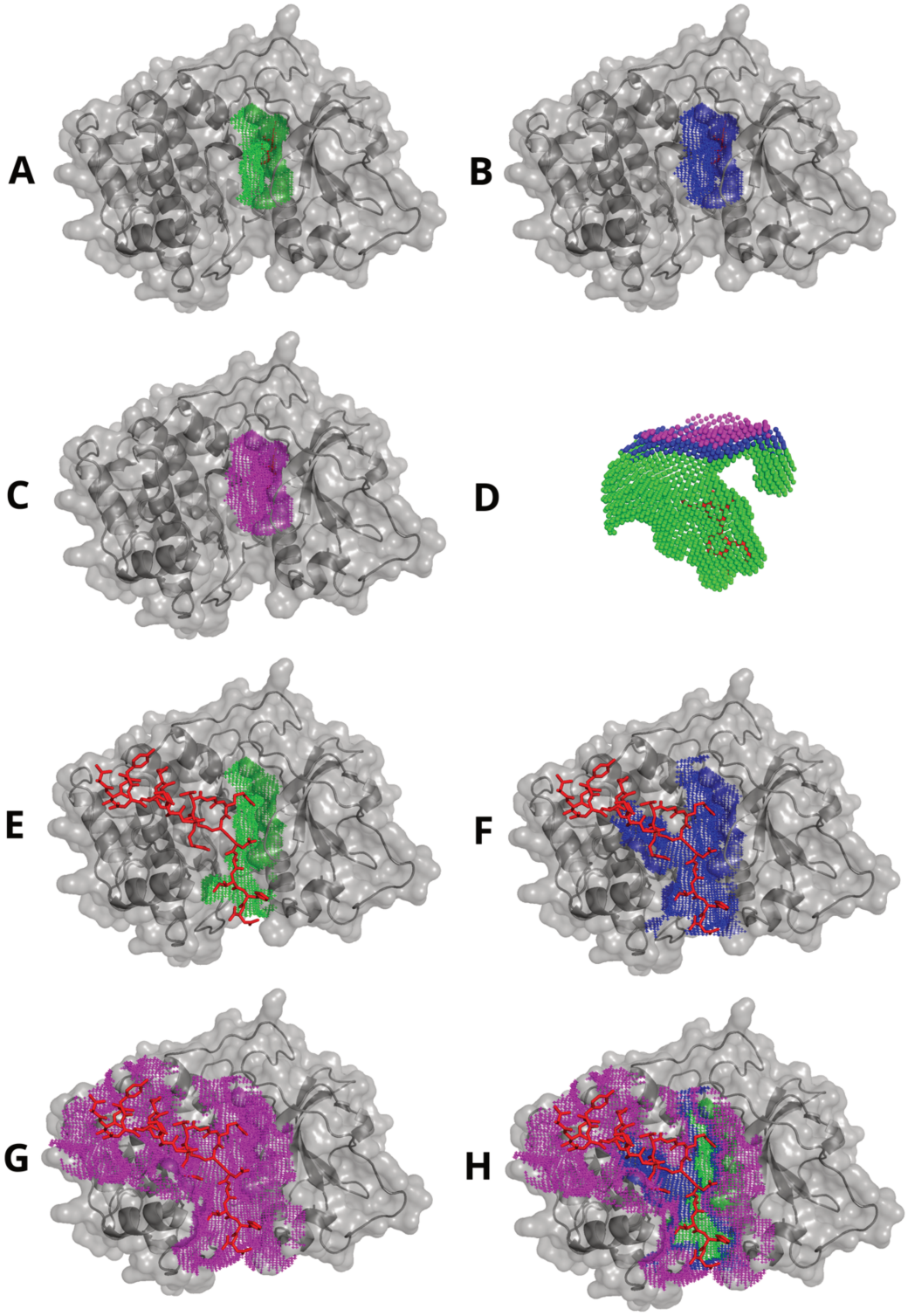


Figura 29: Investigação dos efeitos do parâmetro de distância de remoção. Cavidade do sítio de ligação da adenosina da proteína quinase (PDB 1FMO) para diferentes distâncias de remoção. **(A)** 2.4 ångstroms (verde). **(B)** 1.0 ångstrom (azul). **(C)** 0.0 ångstrom (magenta). **(D)** Sobreposição das cavidades prospectadas do sítio da adenosina com diferentes distâncias de remoção. Cavidade do sítio de ligação da PKI na proteína quinase (PDB 1FMO) para diferentes distâncias de remoção. **(E)** 2.4 ångstroms (verde). **(F)** 1.0 ångstrom (azul). **(G)** 0.0 ångstrom (magenta). **(H)** Sobreposição das cavidades prospectadas do sítio da PKI com diferentes distâncias de remoção.

As prospecções da cavidade do sítio de ligação da adenosina variam com a distância de remoção escolhida, conforme apresentado na Figura 29D. Neste caso, a variação é do teto da cavidade biomolecular, pois a redução da distância de remoção causa a elevação do teto da cavidade. Por outro lado, as detecções da cavidade do sítio de ligação da PKI também variam com o valor de distância de remoção, conforme apresentado na Figura 29H. A variação deste parâmetro também impacta na capacidade de prospectar uma cavidade superficial, sendo que a redução da distância de remoção possibilita a prospecção do sítio de ligação do inibidor PKI, que é um sítio de ligação com baixa profundidade e, portanto, um sítio de ligação superficial. Vale lembrar que a alteração do diâmetro da *Probe Out* também afeta a prospecção de sítios superficiais, assim como o parâmetro "distância de remoção". Por fim, a inclusão do novo parâmetro customizável "distância de remoção" é vinculada a segregação das cavidades prospectadas e a otimização da detecção de cavidades superficiais no novo programa parKVFinder.

5.3.3 Implementação do parâmetro "resolução"

O espaçamento de grade está diretamente associado com a complexidade do sistema a ser resolvido e também à capacidade de prospecção das cavidades biomoleculares. A definição do espaçamento de grade pelo usuário é contra-intuitiva, assim um método de definição indireto aumentaria a usabilidade do programa. O método indireto de definição do espaçamento de grade consiste na escolha de três resoluções: Baixa (*Low*), Média (*Medium*) e Alta (*High*). Estas opções e seus respectivos volumes para os voxels da grade tridimensional estão apresentados na Tabela 3. Vale ressaltar que o método de definição direta do espaçamento de grade continuará disponível por ser útil para usuários experientes do programa, porém apenas um método poderá ser escolhido pelo usuário.

Tabela 3: Volume de voxel definido por cada opção de resolução no programa parKVFinder.

Resolução	Volume do voxel ($Å^3$)
Baixa (*Low*)	0,2
Média (*Medium*)	0,1
Alta (*High*)	0,01

A Figura 30 apresenta a inclusão do método indireto na interface do *plugin* para o PyMOL e a cavidade do sítio de ligação da adenosina na estrutura 1FMO prospectada em diferentes resoluções, mantendo os pontos de cavidade que estão a menos de 8,0 ångstroms da adenosina e usando a sonda *Probe Out* de 8,0 ångstroms de diâmetro. O arquivo de parâmetros com os dois métodos de definição do espaçamento de grade está apresentado no Anexo B.

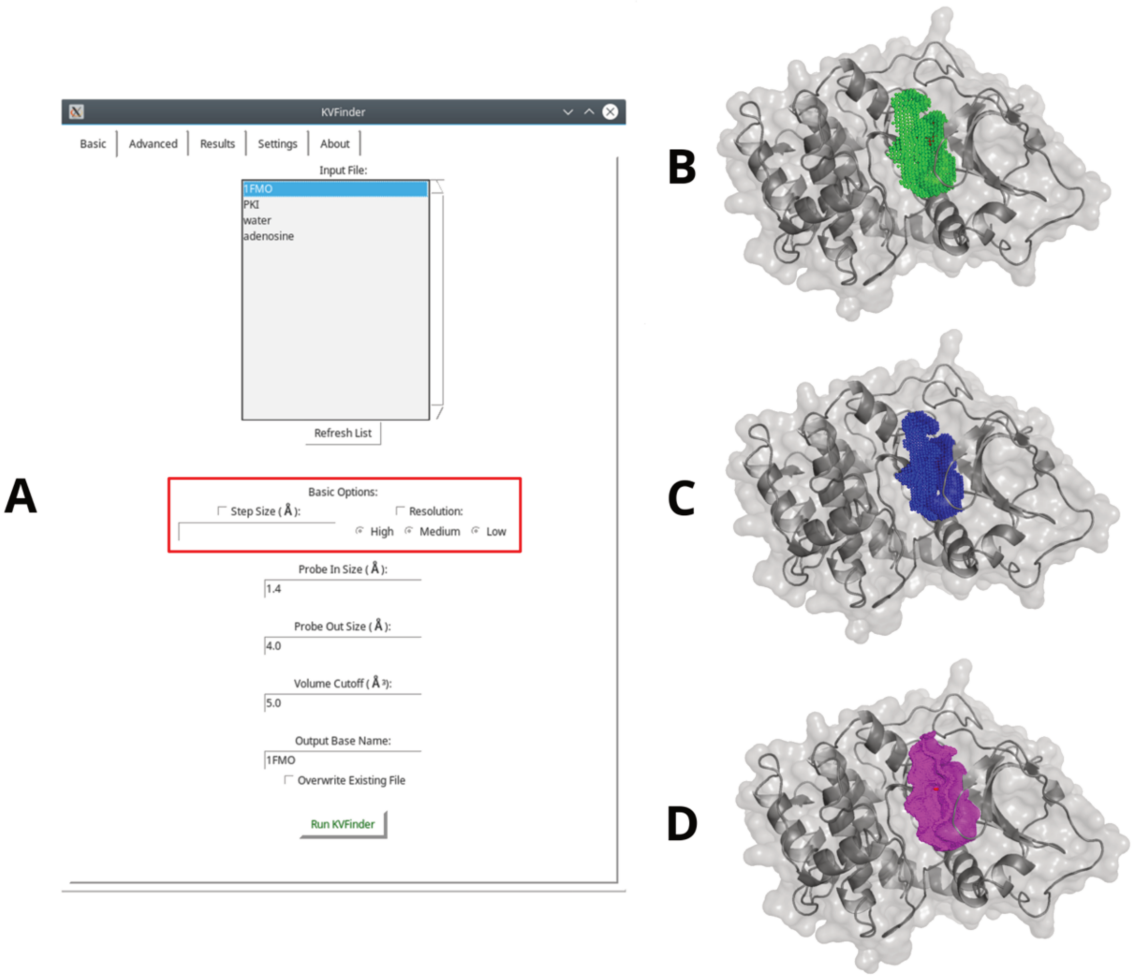


Figura 30: Investigação do novo parâmetro de resolução. (A) Inclusão do parâmetro no *plugin* do PyMOL. As opções do método direto e indireto para definição do espaçamento de grade estão marcadas com um retângulo vermelho. **(B)** Prospecção da cavidade do sítio de ligação da adenosina em resolução Baixa (verde). **(C)** Média (azul). **(D)** Alta (magenta).

O parâmetro “resolução” se baseia no conceito de resolução de imagens digitais da computação gráfica. Esse parâmetro descreve o nível de detalhamento da cavidade biomolecular prospectada. Conforme observado na Figura 30, maiores resoluções significam cavidades mais detalhadas, ou seja, mais voxels descrevendo

um mesmo objeto. Portanto, a inclusão do parâmetro “resolução” simplifica o uso do espaçamento de grade por suas opções serem mais intuitivas aos usuários.

5.3.4 Desenvolvimento da interface de linha de comando

A interatividade com o programa parKVFinder com o usuário final é garantida pelo *plugin* gráfico do PyMOL e pela nova interface de linha de comando integrada ao código-fonte do programa (Figura 31). A interface de linha de comando possibilita a análise de cavidades biomoleculares em larga escala por simplificar chamadas iterativas de arquivos PDBs de entrada e outros parâmetros. As opções de linha de comando podem ser curtas (caractere precedido por ‘-’) e/ou longas (cadeia de caracteres precedido por ‘--’) para definir os parâmetros do programa.

Para a execução do programa parKVFinder através da interface é definido um modo básico de execução, no qual os parâmetros podem ser alterados por meio das opções disponibilizadas, aliada a um conjunto de modos opcionais de execução, os quais podem ser combinados de acordo com a demanda gerada pela estrutura a ser analisada. A configuração básica executa a prospecção e caracterização espacial (volume, área superficial e profundidade) da estrutura completa. Os modos opcionais são: determinação do potencial eletrostático, projeção das escalas de hidrofobicidade, segregação do espaço por ajuste de caixa de busca (caixa de busca customizada e caixa de busca por resíduos), segregação do espaço por ajuste ao ligante e representação da superfície (van der Waals e superfície acessível ao solvente).

A visualização dos arquivos PDB das cavidades (<PDB>*.KVFinder.output.pdb*, <PDB>.<*nome_da_escala*>*.pdb* e <PDB>*.APBS.pdb*) e arquivo de resultados (<PDB>*.KVFinder.results.toml*), permanece compatível com o *plugin* gráfico do PyMOL. Além disso, os arquivos PDB de cavidades podem ser lidos em outros programas de visualização molecular e também podem ser coloridos de acordo com seus valores de fator de temperatura para ilustrar a sua respectiva propriedade. Ainda, o conjunto de parâmetros usados na interface de linha de comando são salvos no formato de arquivo de parâmetros TOML com o intuito de verificar as condições usadas na prospecção e caracterização das cavidades encontradas pelo programa.

Portanto, a interface de linha de comando possibilita a análise em larga escala de cavidades biomoleculares, aumenta a usabilidade do programa e diminui a necessidade da utilização do programa via *plugin* gráfico integrado ao PyMOL.

```
================================================================================
============================== KVFinder help menu ==============================
KVFinder software identifies and describes cavities in target biomolecular
structure using a dual probe system.

The description includes spatial, constitutional and physicochemical
characterization. Spatial characterization includes shape, volume, area, depth
and volume depth. Constitutional characterization includes amino acids type and
class distribution. Physicochemical characterization includes hydrophobicity
scales and electrostatic potential.

Usage: KVFinder PDB [options], where PDB is a path to a target PDB file and
options are:

Options:
   -h, --help            Show this help message and exit.

   -v, --version         Show KVFinder version number and exit.

   --verbose             Print extra information to stdout.

General options:
   -p, --parameters      Define path to parameters file.

   -d, --dictionary      Define path to dictionary file.
                         Default: /home/<user>/KVFinder/dictionary

   -r, --resolution      Define resolution mode (Off, Low, Medium, High)
                         Default: Low

   -s, --step            Define step size (grid spacing).
                         Default: 0.0 A

   -i, --probe_in        Define probe in size.
                         Default: 1.4 A

   -o, --probe_out       Define probe out size.
                         Default: 4.0 A

   --volume_cutoff       Define cavities volume filter.
                         Default: 5.0 A^3

   --removal_distance    Define removal distance when comparing probes
                         surfaces.
                         Default: 2.4 A

   -k, --filled          Output filled cavities. Increase memory consumption
                         for molecular visualization.

   -t, --template        Create a template KVFinder parameters file with
                         default parameters.
                         Default: parameters.toml

   -E, --electrostatic   Calculate electrostatic potential using APBS code.
                         Electrostatic calculations are performed through
                         automatically configured finite difference
                         Poisson-Boltzmann calculations (mg-auto).

   -H, --hydropathy      Map hydrophobicity scales on detected cavities. Native
                         scales includes: HessaHeijne, KyteDoolite,
                         MoonFleming, WimleyWhite and ZhaoLondon.

   -B, --box             Define a search box where KVFinder will detect
                         cavities.

Box adjustment options:
   --custom_box          Define a custom search box based on a file containing
                         the minimum and maximum cartesian values of each axis
                         in angstrom.

   --residues_box        Automatically set a search box based a file containing
                         a tab-separated list of residues.

   --padding             Define residues box padding. Adds a padding length in
                         each box direction.
                         Default: 3.5 A

Surface options:
   -S, --surface         Define a surface representation. The options include:
                         SAS and VdW. SAS specifies solvent accessible surface.
                         VdW specifies van der Waals molecular surface.
                         Default: VdW
```

Figura 31: Menu de ajuda da interface de linha de comando do programa parKVFinder. Os parâmetros podem ser definidos pelos argumentos apresentados no menu. Os parâmetros não definidos usam os valores padrão também apresentados no menu.

```
Ligand options:
   -L, --ligand          Define path to ligand PDB file.

   --ligand_cutoff       Define ligand radius distance cutoff.
                         Default: 5.0 A

Hydropathy option:
   --scales              Define path to hydrophobicity scales dictionary file.
                         Default: /home/<user>/KVFinder/hydrophobicity_scales

Electrostatic options:
   --apbs_results        Load APBS electrostatic potential file in KVFinder.
                         Only accepts the OpenDX multigrid data format.

   --pbe                 Define which Poisson-Boltzmann equation should be
                         solved. The options include npbe, lpbe and lrpbe. npbe
                         specifies nonlinear (full) Poisson-Boltzmann equation.
                         lpbe specifies linearized Poisson-Boltzmann equation.
                         lrpbe specifies linear form of the regularized
                         Poisson-Boltzmann equation.
                         Default: lpbe

   --bcfl                Define type of boundary condition to solve
                         Poisson-Boltzmann equation. The options include: zero,
                         sdh and mdh. zero specifies "zero" boundary condition.
                         sdh specifies "Single Debye-HÜckel" boundary
                         condition. mdh specifies "Multiple Debye-HÜckel"
                         boundary condition.
                         Default: sdh

   --chgm                Define method by which the biomolecular point charges
                         are mapped to the grid for a multigrid Poisson-Boltzm
                         ann equation. The options include: spl0, spl2 and
                         spl4. spl0 specifies traditional trilinear
                         interpolation (linear splines). spl2 specifies cubic
                         B-spline discretization. spl4 specifies quintic
                         B-spline discretization.
                         Default: spl2

   --srfm                Model used to construct the dielectric and
                         ion-accessibility coefficients. The options include:
                         mol, smol, spl2 and spl4. mol specifies dielectric
                         coefficient based on a molecular surface definition
                         and ion-accessibility coefficient based on an
                         "inflated" van der Waals model. smol specifies
                         dielectric and ion-accessibility coefficients as for
                         mol, but both coefficients are "smoothed" by a 9-point
                         harmonic averaging. spl2 specifies dielectric and
                         ion-accessibility coefficients based on a cubic-spline
                         surface. spl4 specifies dielectric and ion-accessibili
                         ty coefficients based on 7th order polynomial.
                         Default: smol

   --pdie                Define solute dielectric constant. This is usually a
                         value between 2 to 20, where lower values consider
                         only electronic polarization and higher values
                         consider additional polarization due to intramolecular
                         motion. The dielectric constant must be greater or
                         equal to 1.
                         Default: 2.0

   --sdie                Define solvent dielectric constant. Bulk water at
                         biologically-relevant temperature is usually modeled
                         with a dielectric constant between 78 to 80. The
                         dielectric constant must be greater or equal to 1.
                         Default: 78.54

   --sdens               Define the number of quadrature points per square
                         angstrom to use in calculation surface terms (e.g.,
                         molecular surface, solvent accessible surface).
                         Default: 10.0 A^-2

   --srad                Define radius of the solvent molecules.
                         Default: 1.4 A

   --swin                Define size of the support for spline-based surface
                         definitions.
                         Default: 0.3 A

   --temp                Define system temperature for calculation.
                         Default: 298.15 K
```

Figura 31 (continuação): Menu de ajuda da interface de linha de comando do programa parKVFinder. Os parâmetros podem ser definidos pelos argumentos apresentados no menu. Os parâmetros não definidos usam os valores padrão também apresentados no menu.

5.4 Descritores de propriedades

Os descritores de propriedades espaciais, constitucionais e físico-químicas estão apresentados e avaliados nos tópicos a seguir. As propriedades espaciais incluem volume, área superficial, forma e profundidade, as propriedades constitucionais incluem composição, características e contagem dos resíduos formadores das cavidades e as propriedades físico-químicas incluem as escalas de hidrofobicidade e o potencial eletrostático.

5.4.1 Descritores de propriedades espaciais

Os descritores de propriedades espaciais, volume, área superficial, forma e profundidade, serão apresentados e avaliados a seguir.

As propriedades espaciais, volume, área superficial e forma, já estavam implementadas no programa KVFinder. No entanto, as mesmas foram atualizadas de forma a aumentar a acurácia. A caracterização espacial melhora com a otimização e paralelização pelo qual o novo programa parKVFinder passou, pois o tempo consumido para um detalhamento maior das cavidades prospectadas foi reduzido. Desta maneira, as mesmas cavidades podem ser prospectadas com um número maior de voxels, gerando uma melhor aproximação de volume, área superficial e forma das cavidades detectadas. Conforme mencionado anteriormente, o filtro de superfície do programa foi atualizado do filtro dos pontos adjacentes para o filtro dos pontos vizinhos, melhorando a caracterização da forma e a estimativa da área superficial.

5.4.1.1 Área superficial

A validação da área superficial de cavidades necessita de um conjunto de sólidos geométricos ocos, uma vez que não é possível verificar a acurácia da área das cavidades prospectadas em estruturas biomoleculares, pois não existe técnica experimental para determinar a área superficial de uma determinada cavidade. Desta forma, o desempenho da nova metodologia de área superficial é avaliado pela comparação da área superficial real e estimada de um conjunto de sólidos geométricos ocos (Figura 32). A comparação das áreas superficiais entre as metodologias de Mullikin e Verbeek (1993), KVFinder e os valores reais de cada sólido estão apresentados na Tabela 4.

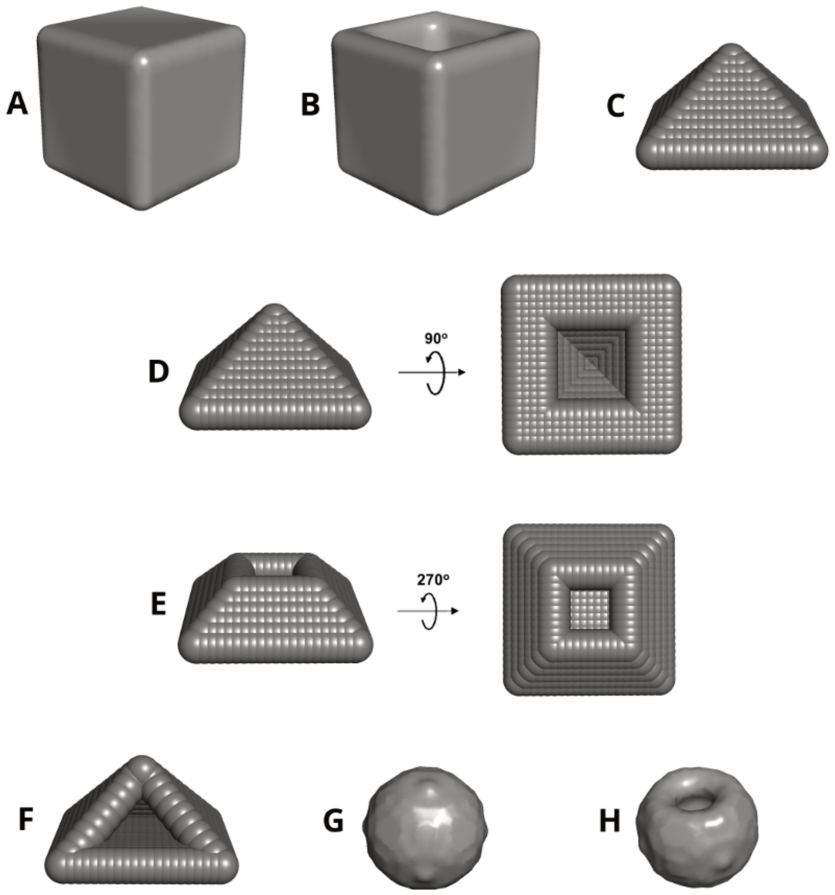


Figura 32: Conjunto de sólidos geométricos ocos.

Tabela 4: Área superficial das cavidades do conjunto de sólidos geométricos ocos. As estimativas da área superficial foram calculadas pelo método do KVFinder e o método adaptado de Mullikin e Verbeek (1993) do parKVFinder.

Sólido	Área real (Å^2)	KVFinder		parKVFinder	
		Área (Å^2)	Erro (%)	Área (Å^2)	Erro (%)
A	360,38	337,62	6,32	311,90	13,45
B	272,25	258,06	5,21	258,15	5,18
C	45,10	32,12	28,78	40,15	10,98
D	21,21	14,06	33,71	19,41	8,49
E	33,98	24,56	27,72	28,29	16,75
F	34,63	26,50	23,48	31,72	8,40
G	84,30	71,62	15,04	83,42	1,04
H	74,22	61,31	17,39	70,63	4,84

Comparando os resultados dos métodos para estimativa da área superficial, observamos que o método adaptado de Mullikin e Verbeek (1993) tem uma redução no erro percentual para todos sólidos geométricos ocos (Tabela 4), exceto para o sólido A (cubo oco). Esse aumento do erro no sólido A ocorre pelo novo método ser punitivo para sistemas que possuem grandes superfícies retas, sendo que o voxel que compõe esse tipo de superfície tem uma face acessível e seu peso é menor que um. Desta maneira, se o objeto tiver grandes extensões de superfície reta, a área superficial será subestimada. Por outro lado, o sólido A tem o erro percentual menor para a metodologia antiga do KVFinder, pois esta metodologia considera as cavidades como objetos compostos por superfícies retas, já que o método classifica todos os voxels de superfície das cavidades com apenas uma face acessível.

No entanto, as cavidades biomoleculares reais são compostas majoritariamente por voxels de superfície, que possuem entre duas e quatro faces acessíveis. Além disso, também não apresentam voxels com apenas uma face acessível de forma consecutiva. Desta maneira, esse sistema para voxels com uma face acessível só se torna punitivo para cavidades quando voxels dessa classe são consecutivos, pois estes voxels consideram que a área superficial é menor que a área da face, pois o voxel não está completamente preenchido pela superfície.

Considerando os sólidos estudados na Figura 32, os sólidos que mais se aproximam do comportamento de uma cavidade biomolecular são os sólidos G (esfera oca) e H (calota esférica oca), pois apresentam uma distribuição de classe de voxels similar à distribuição na cavidade biomolecular. Além disso, os sólidos G e H apresentam o menor erro percentual dentre os sólidos estudados para o novo método implementado no programa parKVFinder.

5.4.1.2 Profundidade

O novo descritor da propriedade espacial de profundidade foi implementado no programa parKVFinder. A metodologia determina a profundidade dos pontos de cavidade, e escreve os valores de profundidade de cada ponto como fator de temperatura no arquivo PDB (Figura 33A) e os valores de profundidade máxima e volume de profundidade da cavidade no arquivo de resultados (Figura 33B). Para facilitar a visualização dos resultados, a interface gráfica do programa PyMOL colore os pontos das cavidades baseado nos valores de profundidade salvos como fator de

temperatura. A paleta de cores utilizada para colorir os pontos é a paleta do arco-íris (Figura 33C), sendo a menor profundidade azul e a maior profundidade vermelho.

A

```
         1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
ATOM     10  HS  KAA   259       2.400   4.200  16.800  1.00  0.00
ATOM     11  HS  KAA   259       2.400   6.000  17.400  1.00  0.00
ATOM     12  HS  KAA   259       2.400   6.600  17.400  1.00  0.00
ATOM     13  HS  KAA   259       2.400   7.200  17.400  1.00  0.00
ATOM     14  HS  KAA   259       2.400   9.000   1.200  1.00 25.94
ATOM     15  HS  KAA   259       2.400   9.000   1.800  1.00 25.06
ATOM     16  HS  KAA   259       2.400   9.600   0.600  1.00 25.69
ATOM     17  HS  KAA   259       2.400   9.600   1.200  1.00 25.79
ATOM     18  HS  KAA   259       2.400   9.600   1.800  1.00 25.08
ATOM     19  HS  KAA   259       2.400   9.600   2.400  1.00 24.12
ATOM     20  HS  KAA   259       2.400  10.200   0.600  1.00 24.92
ATOM     21  HS  KAA   259       2.400  10.200   1.200  1.00 25.02
ATOM     22  HS  KAA   259       2.400  10.200   1.800  1.00 24.82
ATOM     23  HS  KAA   259       2.400  10.200   2.400  1.00 23.85
ATOM     24  HS  KAA   259       2.400  10.800   1.200  1.00 24.27
ATOM     25  HS  KAA   259       2.400  10.800   1.800  1.00 24.41
ATOM     26  HS  KAA   259       3.000   0.600  15.600  1.00  0.00
ATOM     27  HS  KAA   259       3.000   1.200  15.000  1.00  1.00
ATOM     28  H   KAA   259       3.000   1.200  15.600  1.00  0.00
```

B

```
[RESULTS]
#Volume, area, depth and interface residues information for each cavity

      [RESULTS.VOLUME]
      #Volume unit is cubic angstrom
      KAA = 16192.87

      [RESULTS.AREA]
      #Area unit is square angstrom
      KAA = 11456.99

      [RESULTS.DEPTH]
      #Maximum depth unit is angstrom
      KAA = 13.21

      [RESULTS.VOLUME_DEPTH]
      #Volume depth unit is angstrom
      KAA = 133194.48
```

C



Figura 33: Resultados da metodologia de determinação da profundidade das cavidades biomoleculares. (A) Fragmento do arquivo PDB com a coluna de valores de profundidade identificada pelo retângulo vermelho. **(B)** Fragmento do arquivo de resultados com os resultados de profundidade máxima e volume de profundidade das cavidades, identificados pelo retângulo vermelho e azul, respectivamente. **(C)** Paleta de arco-íris para colorir os pontos de cavidade baseado na profundidade.

A profundidade das cavidades pode indicar possíveis sítios de ligação de uma biomolécula, sendo que os sítios de ligação comumente se localizam em regiões onde existe uma variação de profundidade com a vizinhança. As cavidades da estrutura 1FMO foram novamente prospectadas, zerando a distância de remoção para prospectar todas as cavidades proteicas sem segregá-las. A prospecção foi realizada com o diâmetro da sonda *Probe Out* de 10,0 ångströms e espaçamento de grade de

0,6 ångström. As cavidades prospectadas coloridas por profundidade na paleta de arco-íris estão apresentadas na Figura 34.

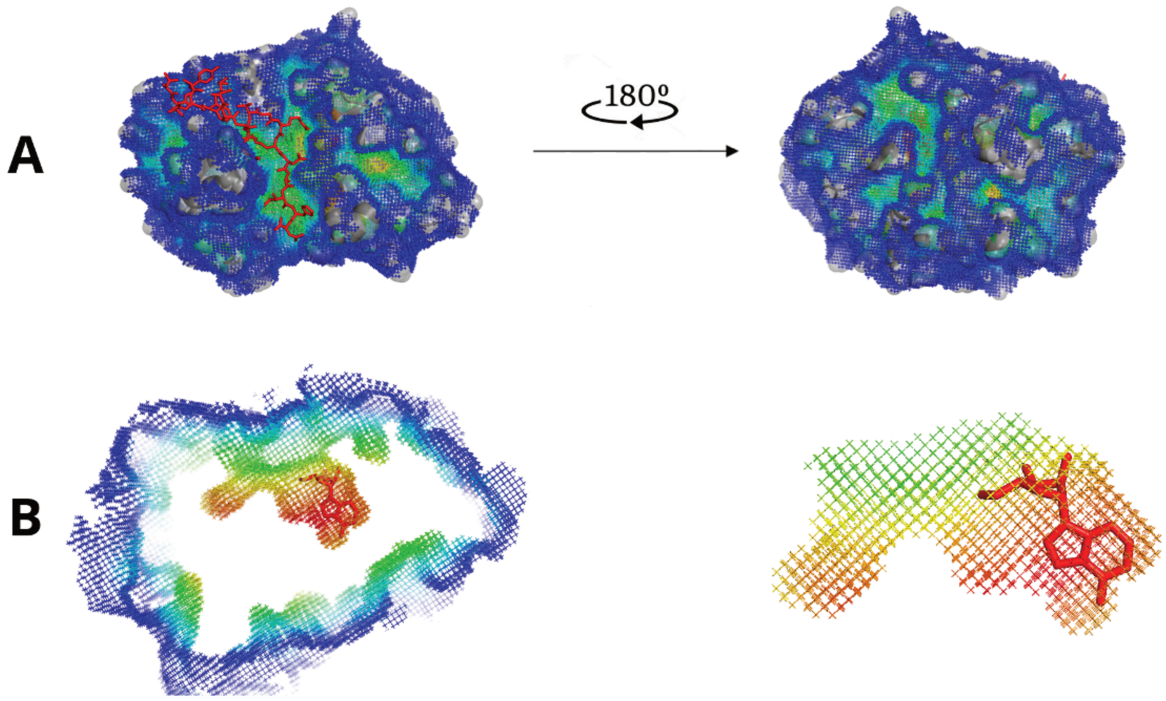


Figura 34: Determinação da profundidade dos sítios de ligação da quinase A. Resultados da prospecção da estrutura da quinase A (PDB 1FMO) com os pontos de cavidade coloridos baseado na profundidade de cada ponto. **(A)** Visão geral das cavidades da proteína quinase com a PKI como *sticks* em vermelho. **(B)** Visão do sítio da adenosina comparado com a superfície (quadro esquerdo) e com foco apenas no sítio (quadro direito).

Baseado na Figura 34, o sítio de ligação da adenosina é o sítio mais profundo da proteína, pois a coloração dos pontos está arrastada para o vermelho. Por outro lado, o sítio de ligação da PKI é menos profundo que o sítio da adenosina, mas o sítio apresenta localmente uma variação da sua profundidade comparada com a vizinhança e a superfície como um todo. Neste caso, o pequeno ligante se liga na cavidade mais profunda da proteína quinase e o peptídeo se liga em um sítio mais superficial que apresenta uma complementaridade geométrica entre a proteína e o peptídeo. Portanto, a busca por cavidades biomoleculares com alta profundidade ou com variação local de profundidade podem ser caminhos interessantes para a identificação de possíveis sítios de ligação.

5.4.2 Descritores de propriedades constitucionais

Os descritores de propriedades constitucionais, composição, características e contagem dos resíduos, serão apresentados e avaliados a seguir.

5.4.2.1 Recuperação dos resíduos formadores das cavidades

A metodologia de recuperação dos resíduos formadores das cavidades foi aprimorada, sendo que o espaço de busca foi alterado de um cubo para uma esfera, conforme ilustrado na Figura 16. A cavidade do sítio de ligação da adenosina da estrutura 1FMO foi novamente prospectada, para os dois espaços de busca, considerando a distância de remoção de 2,4 ångströms para segregar o sítio das demais cavidades. A prospecção foi realizada com o diâmetro da sonda *Probe Out* de 4,0 ångströms e espaçamento de grade de 0,6 ångström. Os resíduos formadores da cavidade do sítio de ligação da adenosina para os dois espaços de busca, cubo e esfera, estão apresentados na Figura 35.

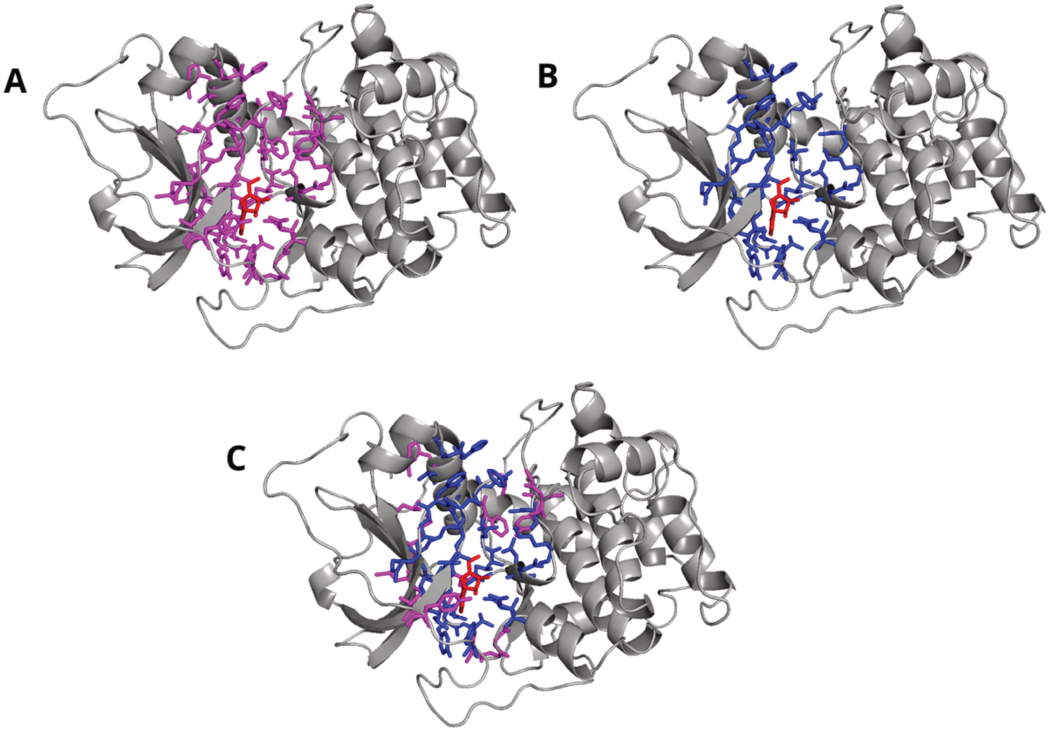


Figura 35: Determinação dos resíduos formadores do sítio de ligação da adenosina da proteína quinase A. Resultados da recuperação dos resíduos formadores do sítio da adenosina (vermelho) na estrutura da quinase A (PDB 1FMO; cinza) para dois diferentes espaços de busca. **(A)** Resíduos recuperados pelo cubo de busca (magenta). **(B)** Resíduos recuperados pela esfera de busca (azul). **(C)** Sobreposição dos resíduos.

Analisando a Figura 35, os resíduos recuperados pela esfera de busca são mais próximos da cavidade encontrada quando comparados aos resíduos recuperados pelo cubo de busca, assim possivelmente reduzindo o número de resíduos que não participam da formação da cavidade. Portanto, a esfera de busca é mais apropriada para a recuperação dos resíduos formadores das cavidades encontradas.

5.4.2.2 Composição, características e contagem dos resíduos formadores

Em posse dos resíduos formadores, a composição das cavidades em termos de tipos de resíduos é conhecida e apresentada no arquivo de resultados (<PDB>*.KVFinder.results.toml*). No entanto, os resíduos ainda podem ser identificados em suas classes pelo esquema de cores no PyMOL, conforme apresentado na Figura 17. Além disso, uma ferramenta foi desenvolvida em Python para a leitura dos arquivos de resultados do programa parKVFinder e criar os histogramas, usando a biblioteca *matplotlib*, da contagem dos tipos e classes de resíduos presentes nas cavidades encontradas. Essa ferramenta será disponibilizada aos usuários junto com o código-fonte do parKVFinder e o *plugin* gráfico do PyMOL. Os resíduos coloridos por classe e os histogramas de tipo e classe de resíduos na cavidade do sítio da adenosina serão apresentados na Figura 36. Os histogramas das cavidades encontradas na estrutura 1FMO são apresentados no Anexo D.

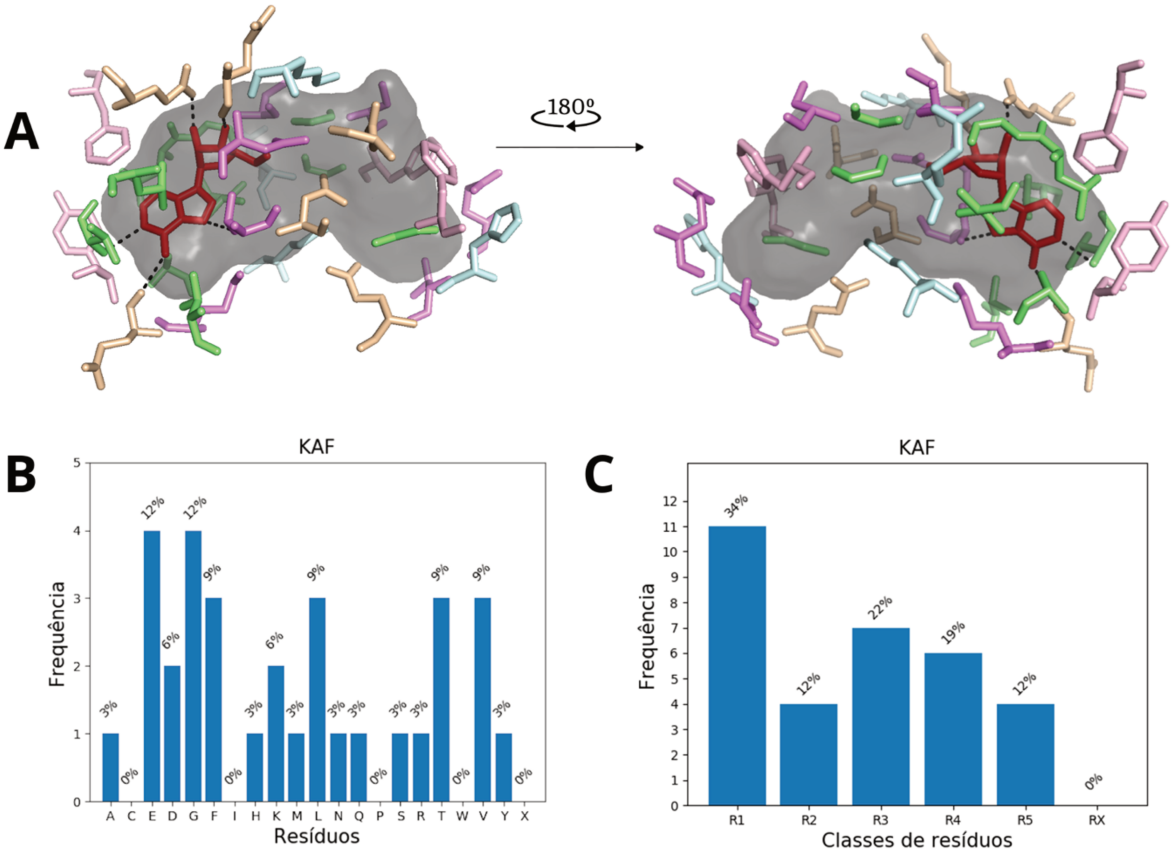


Figura 36: Composição, características e contagem dos resíduos formadores da cavidade do sítio de ligação da adenosina. (A) Resíduos coloridos por classe. R1: verde limão (*lime*). R2: rosa claro (*lightpink*). R3: violeta (*violet*). R4: trigo (*wheat*). R5: ciano pálido (*palecyan*). Linhas tracejadas pretas indicam ligações de hidrogênio, pontes salinas e interações metálicas. **(B)** Histograma da contagem de tipos de resíduos. **(C)** Histrograma da contagem das classes de resíduos.

Analisando a Figura 36A, é possível observar a predominância de resíduos não-polares e alifáticos, classe R1 (verde limão), em torno da porção do sítio onde a adenosina se liga. Os resíduos aromáticos, classe R2 (rosa claro), parecem atuar no fechamento do sítio de ligação, sendo que em ambos lados do sítio os resíduos se combinam em pares para limitar a região do sítio de ligação. Resíduos polares carregados negativamente e não-carregados, classes R3 (violeta) e R4 (trigo), assumem posições estratégicas na proteína para estabilizar o ligante no sítio de ligação através de regiões mais polares. Também existe a contribuição de regiões polares de resíduos apolares, classes R1 (verde limão) e R3 (violeta), para a estabilidade do ligante através de regiões mais apolares. Os resíduos carregados positivamente, classe R5 (ciano pálido), se concentram na região central do sítio de ligação, próxima a região do fosfato-gama removido da adenosina.

Baseado nos histogramas da Figura 36B, a contagem dos tipos de resíduos é homogênea, não apresentando a preferência por nenhum resíduo específico. Já o histograma da Figura 36C, apresenta uma frequência alta de resíduos e uma alta frequência de resíduos polares carregados e não-carregados quando comparado às demais cavidades (Anexo D). Ainda, o sítio apresenta uma preferência por resíduos não-polares e alifáticos, como observado na Figura 36A.

Portanto, os descritores constitucionais do parKVFinder são capazes de fornecer informações interessantes sobre a cavidade do sítio de ligação da adenosina, auxiliando na interpretação do processo de interação entre a proteína e o ligante. Além disso, esses descritores propiciam evidências interessantes sobre quais cavidades são possíveis sítios de ligação. No entanto, vale ressaltar que os constitucionais são apropriados para a caracterização de sítios de ligação em proteínas, sendo pouco informativas para outras biomoléculas, como RNA e DNA.

5.4.3 Descritores de propriedades físico-químicas

Os descritores de propriedades físico-químicas, as escalas de hidrofobicidade e o potencial eletrostático, serão apresentados e avaliados a seguir.

5.4.3.1 Escalas de hidrofobicidade

O descritor de hidrofobicidade foi implementado na forma de escalas de hidrofobicidade projetadas nos pontos de superfície das cavidades proteicas. Porém, existem diversas escalas de hidrofobicidade oriundas de variados métodos de

obtenção e parametrização. Como não existe um consenso a respeito da melhor escala para representar o efeito hidrofóbico, o dicionário de hidrofobicidade foi criado com cinco escalas de hidrofobicidade nativas (Kyte & Doolittle, Hessa & Heijne, Moon & Fleming, Wimley & White e Zhao & London) e a possibilidade de personalização das escalas por parte do usuário final. O dicionário de hidrofobicidade (Figura 18) tem suas informações padronizadas e seu processamento possibilita a inclusão e exclusão de escalas pelo usuário. Vale ressaltar que o descritor de escalas de hidrofobicidade é exclusivo para a caracterização de sítios de ligação em proteínas, não sendo aplicáveis para outras biomoléculas, como RNA e DNA.

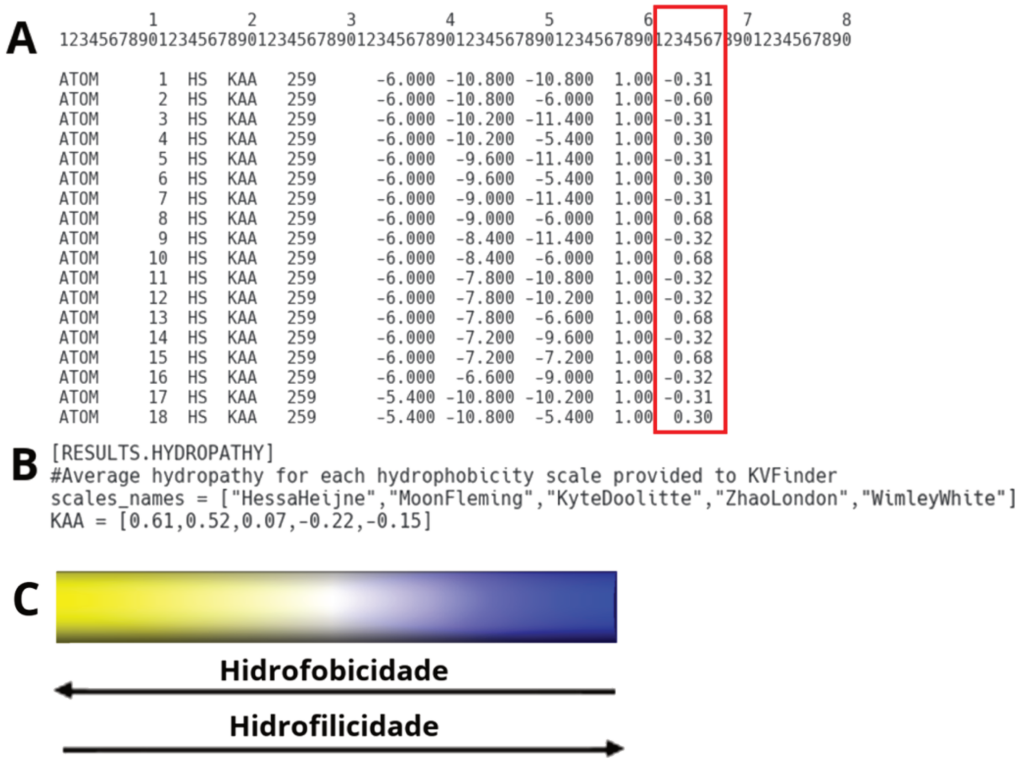


Figura 37: Resultados da metodologia das escalas de hidrofobicidade nas cavidades proteicas. (A) Fragmento do arquivo PDB de saída com a coluna de valores de escala identificada pelo retângulo vermelho. **(B)** Fragmento do arquivo de resultados com os valores da hidropatia média para cada escala nativa do parKVFinder. **(C)** Paleta de amarelo-branco-azul para colorir os pontos de cavidade baseado na escala de hidrofobicidade escolhida.

A metodologia determina o valor de escala aos pontos de superfície das cavidades proteicas, e escreve os valores de escala de cada ponto como fator de temperatura no arquivo PDB (Figura 37A) e os valores médios de hidropatia de cada cavidade no arquivo de resultados (Figura 37B). Para a visualização das escalas no PyMOL, a interface gráfica colore os pontos de superfície das cavidades baseado nos valores de escala salvos como fator de temperatura. A paleta de cores utilizada para

colorir os pontos é a paleta do amarelo-branco-azul (Figura 37C), sendo o valor para o resíduo mais hidrofóbico da escala em amarelo e o valor para o resíduo mais hidrofílico em azul, pois as escalas de hidrofobicidade não tem necessariamente seu valor mínimo representando o resíduo mais hidrofóbico e o valor máximo representando o resíduo mais hidrofílico.

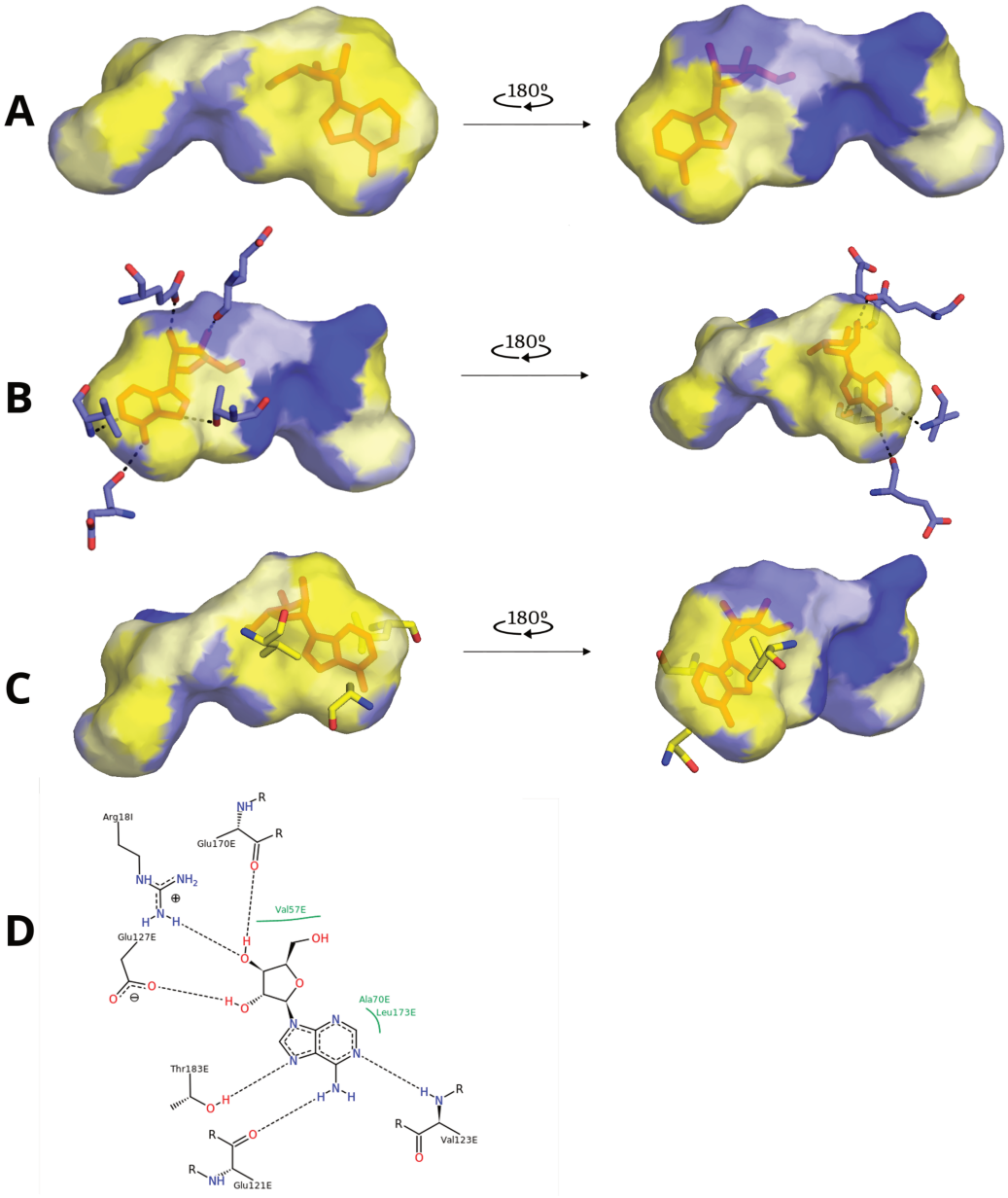


Figura 38: Determinação da hidropatia do sítio de ligação da adenosina da quinase A. Resultados da prospecção do sítio da adenosina na estrutura da quinase A (PDB 1FMO) com os pontos de superfície coloridos para a escala de hidrofobicidade de Hessa & Heijne na paleta amarelo-branco-azul. **(A)** Pontos de superfície da cavidade do sítio de ligação da adenosina, na forma de *surface* com a adenosina na forma de *sticks* (vermelho). **(B)** Interações hidrofílicas entre os resíduos (roxo) e a adenosina (vermelho). **(C)** Interações hidrofóbicas entre os resíduos (amarelo) e a adenosina (vermelho). **(D)** Interações entre os resíduos do sítio de ligação e a adenosina retirado do RCSB PDB. Linhas tracejadas pretas indicam ligações de hidrogênio, pontes salinas e interações metálicas. Linhas verdes mostram interações hidrofóbicas.

A hidropatia das cavidades proteicas pode indicar possíveis tipos de interações nos sítios de ligação de uma proteína. A cavidade do sítio de ligação da adenosina da estrutura 1FMO foi novamente prospectada, considerando a distância de remoção de 2,4 ångströms para segregar o sítio das demais cavidades. A prospecção foi realizada com o diâmetro da sonda *Probe Out* de 4,0 ångströms e espaçamento de grade de 0,25 ångström. A cavidade do sítio de ligação da adenosina colorida pela escala de hidrofobicidade de Hessa & Heijne na paleta de amarelo-branco-azul estão apresentados na Figura 38.

Observando a Figura 38, as escalas de hidrofobicidade possibilitam a identificação dos tipos de interações no sítio de ligação da adenosina na estrutura 1FMO, pois as interações hidrofílicas ocorrem preferencialmente em regiões onde a coloração varia entre branco e azul e as interações hidrofóbicas ocorrem em regiões onde a coloração varia entre amarelo e branco. Vale ressaltar que a estrutura 1FMO não possui os hidrogênios e, com isso, as interações podem estar deslocadas em alguns ångströms, já que as interações foram marcadas entre o átomo no qual o hidrogênio estaria ligado e o átomo em que o hidrogênio faria a ligação de hidrogênio. A interpretação das escalas de hidrofobicidade poderia ser aprimorada com a análise integrada com os descritores constitucionais, assim melhorando a interpretação das interações no sítio de ligação.

5.4.3.2 Potencial eletrostático

O descritor calcula o potencial eletrostático com os pacotes APBS-PDB2PQR, integrados ao programa parKVFinder, com execução assíncrona. Para avaliação e validação do descritor, as cavidades da estrutura 1FMO foram prospectadas com distância de remoção de 2,4 ångströms para segregação do sítio de ligação da adenosina, diâmetro da sonda *Probe Out* de 4,0 ångströms e espaçamento de grade de 0,6 ångström.

O potencial eletrostático é apresentado como linhas de campo (Figura 39A) e superfície molecular (Figura 39B) do mapa eletrostático em formato de dados escalares OpenDX, sendo que esse mapa eletrostático também é projetado nas cavidades encontradas no fator de temperatura do arquivo PDB das cavidades (Figura 39C), e os valores médios de potencial eletrostático de cada cavidade encontrada no arquivo de resultados (Figura 39D). Vale ressaltar que a superfície molecular (Figura 39B) das cavidades é colorido com base nos pontos de cavidade em contato com o

mapa eletrostático, usando as funções nativas do programa PyMOL. Os objetos, linhas de campo, superfície molecular e cavidades, são coloridas pelo potencial eletrostático na paleta de vermelho-branco-azul, usando o intervalo de -1 a 1 kT/e.

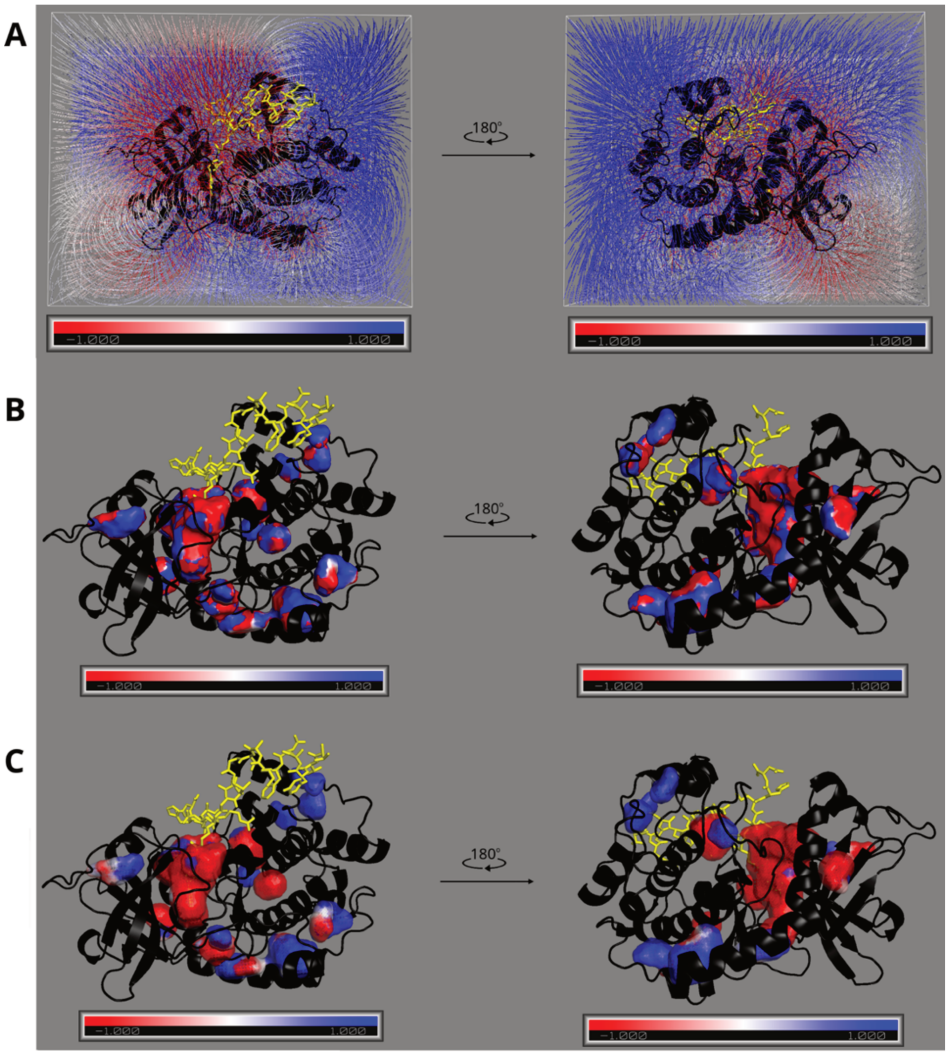


Figura 39: Determinação do potencial eletrostático dos sítios de ligação da quinase A. Resultados do descritor de potencial eletrostático da estrutura da quinase A (PDB 1FMO; preto). Os ligantes, adenosina e PKI, estão representados na forma de *sticks* em amarelo. Os objetos são coloridos na paleta de vermelho-branco-azul, usando o intervalo de -1 a 1 kT/e. **(A)** Linhas de campo do potencial eletrostático. **(B)** Superfície molecular das cavidades encontradas coloridas por potencial eletrostático no PyMOL. **(C)** Cavidades encontradas coloridas pelo potencial eletrostático no fator de temperatura. **(D)** Fragmento do arquivo de resultados com os potenciais médios de cada cavidade encontrada na prospecção. O potencial eletrostático médio da cavidade do sítio da adenosina está identificado com o retângulo vermelho.

Apesar da representação das linhas de campo (Figura 39A) e superfície molecular (Figura 39B) do mapa eletrostático serem interessantes, a projeção do mapa eletrostático nas cavidades encontradas não atinge a mesma acurácia na representação do mapa. Desta forma, a projeção do mapa eletrostático pode ser aprimorada através de métodos de interpolação na inserção dos valores do mapa eletrostático na grade tridimensional, assim possivelmente melhorando a representação do mapa eletrostático nas cavidades, possíveis sítios de ligação, e acurácia do potencial médio calculado no parKVFinder.

O potencial eletrostático é uma forma de representação dos componentes da energética molecular, solvatação e interações eletrostáticas, as quais são relevantes no processo de reconhecimento molecular devido ao grande alcance dessas interações. Baseado na Figura 39, o potencial eletrostático da cavidade do sítio de ligação da adenosina é predominantemente negativo em relação às demais cavidades. Além disso, existe uma diferença de potencial na superfície da proteína, sendo que as linhas de campo (Figura 39A) mostram uma região negativa nas redondezas do sítio de ligação da adenosina e uma outra região positiva na região anterior ao sítio. Portanto, o descritor de potencial eletrostático é capaz de introduzir as características de interações de grande alcance na caracterização das cavidades encontradas.

Por fim, um conjunto de descritores de propriedades espaciais, constitucionais e físico-químicas foram implementados no novo programa parKVFinder para a caracterização das cavidades encontradas na etapa de prospecção. Os descritores espaciais, volume, forma e área superficial, e constitucionais, composição, características e contagem dos resíduos formadores das cavidades, foram aprimorados da versão do KVFinder publicada por Oliveira et al. (2014). Ainda, novos descritores espaciais, profundidade, e físico-químicos, escalas de hidrofobicidade e potencial eletrostático foram implementados no parKVFinder. Essa gama de descritores foi aplicada na estrutura 1FMO como prova de conceito, ilustrando a aplicabilidade e eficiência da caracterização das cavidades, sendo que esses descritores avaliam propriedades importantes para o processo de reconhecimento molecular entre uma biomolécula e um ligante, que pode ser desde uma pequena molécula até um fragmento de DNA. Portanto, o novo programa parKVFinder se torna uma ferramenta interessante para prospecção, caracterização e análise de sítios de ligação em qualquer escala de biomoléculas.

6 CONCLUSÕES E PERSPECTIVAS

O novo programa parKVFinder foi atualizado, otimizado e implementado com rotinas de computação paralela, sendo capaz de prospectar e descrever cavidades em estruturas biomoleculares em diferentes escalas molecular, desde pequenas entidades proteicas até complexos supramoleculares. O novo programa foi compilado em ambiente UNIX, Windows e macOS, apresentando compatibilidade com os sistemas operacionais Ubuntu 18.04, Windows 10 e macOS Mojave. A interatividade com o usuário final é garantida pelo KVFinder PyMOL *plugin*, compatível com o programa de visualização molecular PyMOL v1.8, e a interface de linha de comando integrada ao parKVFinder.

O desempenho computacional do novo programa foi avaliado por meio de um conjunto de testes kv1000, que é formado por um subconjunto do RCSB PDB, contemplando 1000 entidades proteicas que se assemelham nas características da população proteica presente no RCSB PDB. Com base nos resultados de execução desse conjunto, o desempenho de execução do programa parKVFinder aumentou 9,5 vezes em média e melhorou a utilização dos recursos computacionais em comparação com o antigo programa KVFinder. Essa melhora no desempenho possibilitou a inclusão e o aprimoramento dos descritores de propriedades espaciais, constitucionais e físico-químicas, e melhorias incrementais das rotinas no programa parKVFinder.

As melhorias incrementais do programa parKVFinder aumentaram as capacidades de prospecção de cavidades biomoleculares. A inclusão do parâmetro “resolução” simplificou o uso do parâmetro “espaçamento de grade” por ser uma opção mais intuitiva ao usuário, sendo que o parâmetro atua como um método indireto de determinação do espaçamento de grade. Além disso, a inclusão do parâmetro “distância de remoção” otimizou a detecção de cavidades superficiais, sendo possível interferir na segregação das cavidades através desse parâmetro. Por fim, a interface de linha de comando aumentou a interatividade e possibilitou a análise de cavidades biomoleculares em larga escala por parte do usuário final.

A descrição das propriedades das cavidades biomoleculares foi melhorada pelo aprimoramento ou inclusão de descritores espaciais, constitucionais e físico-químicos, porém os descritores constitucionais e as escalas de hidrofobicidade são voltados para cavidades proteicas. A descrição de área superficial e forma foram melhoradas

pela redefinição dos pontos de superfície das cavidades através de um novo filtro de pontos vizinhos. Além disso, a metodologia da determinação da área superficial foi aprimorada por um método baseado na conectividade dos pontos de superfície com a biomolécula, que reduziu o erro percentual da área estimada para a área real em um conjunto de sólidos geométricos ocos. A descrição de profundidade se mostra importante para identificar cavidades biomoleculares que são possivelmente sítios de ligação, conforme apresentado para sítios mais profundos como o sítio da adenosina na estrutura 1FMO e para os sítios superficiais como o sítio da PKI na estrutura 1FMO.

A descrição constitucional das cavidades em termos da composição, características e contagem dos resíduos formadores auxiliam na interpretação do processo de reconhecimento molecular entre o receptor e o ligante. A descrição de hidropatia por meio das escalas de hidrofobicidade possibilitou a identificação de características estruturais bem estabelecidas, como ligações de hidrogênio e outros tipos de interações no sítio de ligação da adenosina na estrutura 1FMO. Além disso, a customização do dicionário de hidrofobicidade traz a possibilidade de investigar o problema a partir de diferentes escalas e incluir escalas que não são nativas do programa parKVFinder. Por fim, a descrição da solvatação e interações eletrostáticas por meio do potencial eletrostático calculado pelos pacotes APBS-PDB2PQR, integrados ao parKVFinder, introduzem informações sobre interações de grande alcance na caracterização das cavidades biomoleculares encontradas.

As perspectivas futuras passam pelo desenvolvimento de um preditor de sítios de ligação em estruturas biomoleculares, baseado na caracterização das cavidades biomoleculares prospectadas pelo parKVFinder, sendo que o programa ainda é dependente da interpretação humana para avaliação de quais cavidades são possíveis sítios de ligação. Com a evolução e disseminação do aprendizado profundo, o computador é capaz de reconhecer padrões, algumas vezes imperceptíveis por humanos, sendo possível o reconhecimento de padrões de sítios de ligação em estruturas biomoleculares. Por outro lado, a correlação dos descritores nas cavidades que são sítios de ligação e as que não são deve ser avaliada em um grande conjunto dados para determinar as variáveis (descritores de propriedades) relevantes para o problema estudado. Além disso, outra rodada de aprimoramento dos descritores pode ser necessária, como, por exemplo, a melhoria na projeção do mapa eletrostático nas cavidades por meio de métodos de interpolação.

7 REFERÊNCIAS

AMDAHL, G. M. Validity of the single processor approach to achieving large-scale computing capabilities. *Proceedings of AFIPS Spring Joint Computer Conference,* AFIPS Press*,* p. 483-485, 1967.

AN, J.; TOTROV, M.; ABAGYAN, R. Comprehensive identification of “druggable” protein ligand binding sites. *Genome informatics,* International Conference on Genome Informatics, v. 15, n. 2, p. 31-41, 2004.

AN, J.; TOTROV, M.; ABAGYAN, R. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Molecular and Cellular Proteomics*, American Society for Biochemistry and Molecular Biology, v. 4, n. 6, p. 752-761, 2005.

ARMON, A.; GRAUR, D.; BEN-TAL, N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *Journal of Molecular Biology*, Elsevier BV, v. 307, n. 1, p. 447-463, 2001.

BAKER, N. A. et al. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Science*. Proceedings of the National Academy of Science, v. 98, n. 1, p. 10037-10041, 2001.

BARTLETT, G. J. et al. Analysis of catalytic residues in enzyme active sites. *Journal of Molecular Biology*, Elsevier BV, v. 324, n. 1, p. 105-121, 2002.

BERMAN, H. M et al. The Protein Data Bank. *Nucleic Acids Research*, Oxford University Press (OUP), v. 28, n .1, p. 235-242, 2000.

BISWAS, K. M.; DEVIDO, D. R.; DORSEY, J. G. Evaluation of methods for measuring amino acid hydrophobicities and interactions. *Journal of Chromatography A*, Elsevier BV, v. 1000, n. 1-2, p. 637-655, 2003.

BOHACEK, R. S.; MCMARTIN, C. Modern computational chemistry and drug discovery: structure generating programs. *Current Opinion in Chemical Biology*, Elsevier BV, v. 1, n. 2, p. 157-161, 1997.

BRADFORD, J. R.; WESTHEAD, D. R. Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, Oxford University Press (OUP), v. 21, n. 8, p. 1487-1494, 2005.

BRADY, G. P.; STOUTEN, P. F. Fast prediction and visualization of protein binding pockets with PASS. *Journal of Computer-Aided Molecular Design*, Springer Nature, v. 14, p. 383-401, 2000.

BRYLINSKI, M.; SKOLNICK, J. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proceedings of the National Academy of Sciences*, Proceedings of the National Academy of Sciences, v. 105, p. 129-134, 2008.

CAO, B.; YANG, M.; MAO, C. Phage as a Genetically Modifiable Supramacromolecule in Chemistry, Materials and Medicine. *Accounts of Chemical Research*, American Chemical Society (ACS), v. 49, n. 6, p. 1111-1120, 2016.

CARLSON, H. A. et al. Differences between high- and low-affinity complexes of enzymes and nonenzymes. *Journal of Medicinal Chemistry*. American Chemical Society (ACS), v. 51, n. 20, p. 6432-6441, 2008.

DABERDAKU, S.; FERRARI, C. Computing voxelised representations of macromolecular surfaces: A parallel approach. *International Journal of High Performance Computing Applications*, SAGE Publications, v. 32, n. 3, p. 407-432, 2016.

DODSON, G.; WLODAWER, A. Catalytic triads and their relatives. *Trends in Biochemical Sciences*, Cell Press, v. 23, n. 9, p. 347-352, 1998.

DOLINSKY, T. J.; NIELSEN J. E.; MCCAMMON J. A.; BAKER N. A. PDB2PQR: an automated pipeline for the setup, execution, and analysis of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Research*, Oxford University Press (OUP), v. 32, n. 1, p. W665-W667, 2004.

FOSTER, I. *Designing and Building Parallel Programs: concepts and tool for parallel software engineering*. [S.l]: Pearson, 1995.

GHERSI, D.; SANCHEZ, R. Improving accuracy and efficiency of blind protein-ligand docking by focusing on predicted binding sites. *Proteins*, Wiley-Blackwell, v. 74, n. 2, p. 417-424, 2010.

GOODFORD, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of Medicinal Chemistry*, American Chemical Society (ACS), v. 28, n. 7, p. 849-857, 1985.

GRAMA, A. et al. *Introduction to Parallel Computing*. 2 ed. Harlow: Pearson, 2003.

GUILLOUX, V. L.; SCHMIDTKE, P.; TUFFERY, P. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics*, Springer Nature, v. 10, n. 1, p. 168, 2009.

GUSTAFSON, J. L. Reevaluating Amdahl's Law. *Communications of the ACM*, Association for Computing Machinery, v. 31, n. 5, p. 532-533, 1988.

HALGREN, T. New method for fast and accurate binding-site identification and analysis. *Chemical Biology and Drug Design*, Wiley-Blackwell, v. 69, n. 2, p. 146-148, 2007.

HARRIS, R. C. et al. Numerical difficulties computing electrostatics potentials near interfaces with the poisson-boltzmann equation. *Journal of Chemical Theory and Computation*, American Chemical Society (ACS), v. 13, n. 8, p. 3945-3951, 2017.

HE, Z.; ZHANG, C.; XU, Y.; ZENG, S.; ZHANG, J.; XU, D. MUFOLD-DB: a processed protein structure database for protein structure prediction and analysis. *BMC Genomics*, Springer Nature, v. 15, n. Suppl 11, p. S2, 2014.

HEIDEN, W.; MOECKEL, G.; BRICKMANN, J. A new approach to analysis and display of local lipophilicity/hydrophilicity mapped on molecular surfaces. *Journal of Computer-Aided Molecular Design*, Springer Nature, v. 7, n. 5, p. 503-514, 1993.

HENDLICH, M.; RIPPMANN, F.; BARNICKEL, G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal of Molecular Graphics and Modelling*, Elsevier BV, v. 15, n. 6, p. 359-363, 1997.

HENRICH, S. et al. Computational approaches to identifying and characterizing protein binding sites for ligand design. *Journal of Molecular Recognition*, Wiley-Blackwell, v. 23, n. 2, p. 209-219, 2010.

HESSA, T. et al. Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature*, Springer Nature, v. 433, p. 377-381, 2005.

HO, C. M.; MARSHALL, G. R. Cavity search: an algorithm for the isolation and display of cavity-like binding regions. *Journal of Computer-Aided Molecular Design*, Springer Nature, v. 4, n.4, p. 337-354, 1990.

HONIG, B.; NICHOLLS, A. Classical electrostatics in biology and chemistry. *Science*, American Association of the Advancement of Science, v. 268, n. 5214, p. 1144-1149, 1995.

HUANG, B.; SCHROEDER, M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Structural Biology*, Springer Nature, v. 6, p. 19, 2006.

HUANG, B. MetaPocket: a meta approach to improve protein ligand binding site prediction. *OMICS: A Journal of Integrative Biology*, Mary Ann Liebert, v. 13, n. 4, p. 325-330, 2009.

HUBBARD, S. J.; ARGOS, P. Cavities and packing at protein interfaces. *Protein Science*, WIley-Blackwell, v. 3, n. 12, p. 2194-2206, 1994.

JANA, T et al. PPIMpred: a web server for high-throughput screening of small molecules targeting protein-protein interaction. *Royal Society Open Science*, The Royal Society, v. 4, n. 4, p. 160501, 2017.

JO, S. et al. PBEQ-Solver for online visualization of electrostatics potential of biomolecules. *Nucleic Acid Research*, Oxford University Press (OUP), v. 36, p. W270-W275, 2008.

JUNCKER, A. S. et al. Sequence-based feature prediction and annotation of proteins. *Genome Biology*, Springer Nature, v. 10, n. 2, p. 206, 2009.

JURRUS, E. et al. Improvements to the APBS biomolecular solvatation software suite. *Protein Science*, Wiley-Blackwell, v. 27, n. 1, p. 112-128, 2017

KAWABATA, T.; GO, N. Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites. *Proteins*. Wiley-Blackwell, v. 68, p. 516-529, 2007.

KELLOGG, G. E.; SEMUS, S. F.; ABRAHAM, D. J. HINT: a new method of empirical hydrophobic field of calculation for CoMFA. *Journal of Computer-Aided Molecular Design*, Springer Nature, v. 5, n. 6, p. 545-552, 1991.

KINJO, A. R. et al. Protein Data Bank Japan (PDBj): Maintaining a structural data archive and resource description framework format. *Nucleic Acids Research*, Oxford University Press (OUP), v. 40, p. D453-D460, 2012.

KLEYWEGT, G. J.; JONES, T. A. Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallographica, Section D: Biological Crystallography*, v. 50, p. 178-185, 1994.

KOKH, D. B. et al. TRAPP: a tool for analysis of transient binding pockets in proteins. *Journal of Chemical Information and Modelling*, American Chemical Society (ACS), v. 53, n. 5, p. 1235-1252, 2013.

KUHN, D. et al. From the similarity analysis of protein cavities to the functional classification of protein families using cavbase. *Journal of Molecular Biology*, Elsevier BV, v. 359, p. 1023-1044, 2006.

KYTE, J.; DOOLITTLE, R. F. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, Elsevier BV, v. 157, n. 1, p. 105-132, 1982.

LAMM, G. The poisson-boltzmann equation. *Reviews in Computational Chemistry*. New York: John Wiley & Sons. p. 147-365, 2003.

LASKOWSKI, R. A. et al. Protein clefts in molecular recognition and function. *Protein Science: a publication of the Protein Society*, Elsevier BV, v. 5, n. 12, p. 2438-2452, 1996.

LASKOWSKI, R. A. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of Molecular Graphics and Modelling*, Elsevier BV, v. 13, n. 5, p. 323-330, 1995.

LAURIE, A. T. R.; JACKSON, R. M. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics*, Oxford University Press (OUP), v. 21, n. 9 p. 1908-1916, 2005.

LEHNINGER, A. L.; NELSON, D. L.; COX, M. M. *Princípios de Bioquímica*. 2ed. São Paulo: Sarvier, 1995.

LEVITT, D. G.; BANASZAK, L. J. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *Journal of Molecular Graphics and Modelling*, Elsevier BV, v. 10, n. 4, p. 229-234, 1992.

LIANG, J.; WOODWARD, C.; EDELSBRUNNER, H. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein Science*, Wiley-Blackwell, v. 7, n. 9, p. 1884-1897, 1998.

MADURA, J. D. et al. Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program. *Computer Physics Communications*, Elsevier BV, v. 91, n. 1-3, p. 57-95, 1995.

MANNHOLD, R. et al. Calculation of molecular lipophilicity: state-of-the-art and comparison of log P methods on more than 96,000 compounds. *Journal of Pharmaceutical Sciences*, Elservier BV, v. 98, n. 3, p. 861-893, 2009.

MATHERON, G. *Random sets and integral geometry (Probability and Mathematical Statistics)*. New York: John Wiley & Sons, 1975.

MOON, C. P.; FLEMING, K. G. Side-chain hydrophobicity scale derived from transmembrane protein folding into lipid bilayers. *Proceedings of the National Academy of Sciences of the United States of America*, Proceedings of the National Academy of Sciences, v. 108, n. 25, p. 10174-10177, 2011.

MORRIS, G. M. et al. Autodock4 and AutoDockTools4: automated docking with selective receptor flexiblity. *Journal of Computational Chemistry*, Wiley-Blackwell, v. 30, n. 16, p. 2785-2791, 2009.

MULLIKIN, J. C.; VERBEEK, P. W. Surface area estimation of digitized planes. *Bioimaging*, Wiley-Blackwell, v. 1, n. 1, p. 6-16, 1993.

NAYAL, M.; HONIG B. On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins*, Wiley-Blackwell, v. 63, p. 892-906, 2006.

NETTO, A. V. G.; FREM, R. C. G.; MAURO, A. E. A química supramolecular de complexos pirazólicos. *Química Nova*, Sociedade Brasileira de Química, v. 31, n. 5, p. 1208-1217, 2008.

NICHOLLS, A.; SHARP, K. A.; HONIG, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins*, Wiley-Blackwell, v. 11, n. 4, p. 281-296, 1991.

NUÑEZ-VIVANCO, G. et al. Geomfinder: a multi-feature identifier of similar three-dimensional protein patterns: a ligand-independent approach. *Journal of Cheminformatics*, Springer Nature, v .8, p. 19, 2016.

OBERHAUSER, N.; NURISSO, A.; CARRUPT, P. A. MLP Tools: a PyMOL plugin for using the molecular lipophilicity potential in computer-aided drug design. *Journal of Computer-Aided Molecular Design*, Springer Nature, v. 28, n. 5, p. 587-596, 2014.

OLIVEIRA, S. H. et al. KVFinder: steered identification of protein cavities as a PyMOL plugin. *BMC Bioinformatics*, Springer Nature, v. 15, n. 1, p. 197, 2014.

OLIVEIRA, S. H. P. *Desenvolvimento de um algoritmo para identificação e caracterização de cavidade em regiões específicas de estruturas tridimensionais de proteínas*. Tese (Doutorado), 2011.

OSADA, H. *Protein targeting with small molecules: chemical biology techniques and applications*. Hoboken: John Wiley & Sons, 2009.

PATTERSON, D. A.; HENNESSY, J. L. *Computer Organization and Design*. 3 ed. San Francisco: Morgan Kaufmann Publishers, 2005.

PETREK, M. et al. CAVER: a new tool to explore routes from protein clefts, pockets and cavities. *BMC Bioinformatics*, Springer Nature, v. 7, n. 1, p. 316, 2006.

PIETERS, B. J. G. E. et al. Natural supramolecular protein assemblies. *Chemical Society Reviews*, Royal Society of Chemistry, v. 45, n. 1, p. 24-39, 2016.

PUPKO, T. et al. Rate4site: an algorithmic tool for identification of functional regions in proteins by surface mapping evolutionary determinants within their homologues. *Bioinformatics*, Oxford University Press (OUP), v. 18, n. Suppl 1, p. S71-S77, 2002.

ROSE, P. W. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Research*, Oxford University Press (OUP), v. 45, n. D1, p. D271-D281, 2017.

SCHIMITT, S.; HENDLICH, M.; KLEBE, G. From structure to function: A new approach to detect functional similarity among proteins independent from sequence and fold homology. *Angewandte Chemie International Edition*, Wiley-Blackwell, v. 40, n. 17, p. 3141-3144, 2001.

SCHRÖDINGER, LCC. The PyMOL molecular graphics system, version 1.8. 2015.

SCOTT, D. E. et al. Using a fragment-based approach to target protein-protein interactions. *ChemBioChem*, Wiley-Blackwell, v. 14, n. 3, p. 332-342, 2013.

SERRA, J. *Image analysis and mathematical morphology.* 1 ed. London: Academic Press, 1982.

SILLITOE, I. et al. CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acid Research*, Oxford University Press (OUP), v. 43, n. D1, p. D376-381, 2014.

SMITH, N. et al. Delphi web server v2: incorporating atomic-style geometrical figures into the computational protocol. *Bioinformatics*, Oxford University Press (OUP), v. 28, n. 12, p. 1655-1657, 2012.

SOGA, S. et al. Use of amino acid composition to predict ligand-binding site. *Journal of Chemical Information and Modelling*, American Chemical Society (ACS), v. 47, n. 2, p. 400-406, 2007.

SOTRIFFER, C.; KLEBE, G. Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design. *Farmaco (Societa chimica italiana: 1989)*, v.57, n.3, p. 243-251, 2002.

STANK, A. et al. Protein binding pockets dynamics. Accounts of chemical research, American Chemical Society (ACS), v. 49, n. 5, p. 809-815, 2016.

TARJAN, R. Depth-First Search and Linear Graph Algorithms. *Journal on Computing*, Society for Industrial and Applied Mathematics (SIAM), v. 1, n. 2, p. 146-160, 1972.

TETKO, I. V.; TANCHUK, V. Y. Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. *Journal of Chemical Information and Computer Sciences*, American Chemical Society (ACS), v. 42, n. 5, p. 1136-1145, 2002.

THORNTON, J. M. et al. From structure to function: approaches and limitations. *Nature Structural and Mocelular Biology*, Springer Nature, v. 7, p. 991-994, 2000.

UNNI, S. et al. Web servers and services for electrostatics calculations with apbs and pdb2pqr. *Journal of Computational Chemistry*, Wiley-Blackwell, v. 32, n. 7, p. 1488-1491, 2011.

VELANKAR, S. et al. PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic Acids Research*, Oxford University Press (OUP), v. 44, n. D1, p. D385-D395, 2016.

VOLKAMER, A. et al. Analyzing the topology of active sites: on the prediction of pockets and subpockets. *Journal of Chemical Information and Modelling*, American Chemical Society (ACS), v. 50, n. 11, p. 2041-2052, 2010.

WANG, C. Exploring accurate poisson-boltzmann methods for biomolecular simulations. *Computational and Theoretical Chemistry*, Elsevier BV, v. 1024, p. 34-44, 2013.

WEBBER, M. J. et al. Supramolecular biomaterials. *Nature Materials*, Springer Nature, v. 15, p. 13-26, 2016.

WEISEL, M.; PROSCHAK, E.; SCHNEIDER, G. PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chemical Central Journal*, Springer Nature, v. 1, n. 1, p.7, 2007.

WIMLEY, W. C.; WHITE, S. H. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nature Structural Biology*, Springer Nature, v. 3, n. 10, p. 842-848, 1996.

WINDREICH, G.; KIRYATI, N.; LOHMANN, G. Voxel-based surface area estimation: from theory to practice. *Pattern Recognition*, Elservier BV, v. 36, p. 2531-2541, 2003.

XIE, Y.; YING, J.; XIE, D. SMPBS: web server for computing biomolecular electrostatics using finite element solvers of size modified poisson-boltzmann equation. *Journal of Computational Chemistry*, Wiley-Blackwell, v. 38, n. 8, p. 541-552, 2017.

YOSHII, T. et al. Intracellular protein-responsive supramolecules: Protein sensing and in-cell construction of inhibitor assay system. *Journal of the American Chemical Society*, American Chemical Society (ACS), v. 136, n. 47, p. 16635-16642, 2014.

YU, J. et al. Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. *Bioinformatics*, Oxford University Press (OUP), v. 26, n. 1, p. 46-52, 2010.

ZHAO, G.; LONDON, E. An amino acid "transmembrane tendency" scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: relationship to biological hydrophobicity. *Protein Science*, Wiley-Blackwell, v. 15, n. 8, p. 1987-2001, 2006.

ZHU, H. et al. A new group contribution approach to the calculation of log P. *Current Computer-Aided Drug Design*, Bentham Science, v. 1, n. 1, p. 3-9, 2005.

ZITVOGEL, L. et al. Immunological off-target effects of imatinib. *Nature Review Clinical Oncology*, Springer Nature, v. 13, n. 7, p. 431-446, 2016.

ANEXO A: EXEMPLOS DE ARQUIVOS DE CONFIGURAÇÃO

Anexo A1: Exemplo de arquivo de configuração no formato TOML.

```toml
# This is a TOML document.

title = "TOML Example"

[owner]
name = "João Victor da Silva Guerra"
organization = "Brazilian Biosciences National Laboratory/CNPEM"
group = "Laboratory of Computational Biology"

[academic]
project = "Prospection and characterization of supramolecular cavities"
supervisor = "Paulo Sergio Lopes de Oliveira"
level = "master's scholar"
year = 2018
program = "Biosciences and Technology of Bioactive Products"
institute = "Institute of Biology"
university = "University of Campinas"
```

Anexo A2: Exemplo de arquivo de configuração no formato JSON.

```json
{
  "title": "JSON Example",
  "owner": {
    "name": "João Victor da Silva Guerra",
    "organization": "Brazilian Biosciences National Laboratory/CNPEM",
    "group": "Laboratory of Computational Biology"
  },
  "academic": {
    "project": "Prospection and characterization of supramolecular cavities",
    "supervisor": "Paulo Sergio Lopes de Oliveira",
    "level": "master's scholar",
    "year": 2018,
    "program": "Biosciences and Technology of Bioactive Products",
    "institute": "Institute of Biology",
    "university": "University of Campinas"
  }
}
```

Anexo A3: Exemplo de arquivo de configuração no formato YAML.

```yaml
# This is an YAML document.
title: YAML Example
owner:
  name: João Victor da Silva Guerra
  organization: Brazilian Biosciences National Laboratory/CNPEM
  group: Laboratory of Computational Biology
academic:
  project: Prospection and characterization of supramolecular cavities
  supervisor: Paulo Sergio Lopes de Oliveira
  level: master's scholar
  year: 2018
  program: Biosciences and Technology of Bioactive Products
  institute: Institute of Biology
  university: University of Campinas
```

ANEXO B: ARQUIVOS DO PROGRAMA KVFINDER

Anexo B1: Exemplo de arquivo de parâmetros Parameters.txt (versão antiga) do programa KVFinder.

```
/home/user/KVFinder/dictionary
input/1FMO.pdb
1FMO.KVFinder.output
1
0.6
0
0 0 0
0 0 0
0 0 0
0 0 0
0 0 0
0 0 0
0 0 0
0 0 0
1.4
4.0
0
5.0
1
0
0
-
0.0
```

Anexo B2: Exemplo de arquivo de parâmetros *parameters.toml* (versão nova) do programa parKVFinder.

```
#TOML configuration file for KVFinder software.

title = "KVFinder parameters file"

[FILES_PATH]
#Path for the KVFinder dictionary file, with atoms information.
dictionary = "/home/user/KVFinder/dictionary"
#Path for the hydrophobicity scales dictionary file.
hydropathy_dictionary = "/home/user/KVFinder/hydrophobicity_scales"
#Path for the input PDB file.
pdb = "/home/user/KVFinder/input/1FM0.pdb"
#Path for the output files.
output = "/home/user/KVFinder/input/"
#Base name for the output files.
base_name = "1FM0"
#Path for the PDB ligand file.
ligand = ""

[SETTINGS]
#Settings for KVFinder software

    [SETTINGS.modes]
    #Whole Protein mode defines the search space as the whole protein.
    whole_protein_mode = true
    #Box Ajustment mode defines the search space as the box drawn in PyMOL.
    box_mode = false
    #Resolution mode automatically defines step size inside the box grid.
    #If set High, it defines a unitary box volume of 0.2. If set Medium, it defines a unitary box volume of 0.1. If set Low, it defines a
    unitary box of 0.01. If set Off, it gets inputed step size.
    resolution_mode = "Low"
    #Surface mode selects type to be considered, (false) Solvent Accessible Surface (SAS) or (true) Molecular Surface (VdW).
    surface_mode = true
    #kvp_mode defines if cavities output PDB will be shown as filled cavities (true) or not (false).
    kvp_mode = false
    #Ligand mode is used to limit the search space around a ligand.
    ligand_mode = false
    #Hydrophathy mode maps hydrophobicity scales on the prospected cavities.
    hydropathy_mode = true
    #Electrostatic mode activates calculation electrostatic potential through APBS code.
    electrostatic_mode = true

    [SETTINGS.step_size]
    #Defines the size between grid points. Directly affects the precision. Also has a effect on the running time.
    step_size = 0.00

    [SETTINGS.probes]
    #KVFinder works with a two sized probe system. A smaller probe, the Probe In, and a bigger one, the Probe Out, rolls around the protei
    #The points reached by the Probe In but not by the Probe Out are considered cavity points.
    #The Probe In radius.
    probe_in = 1.40
    #The Probe Out radius
    probe_out = 4.00

    [SETTINGS.cutoffs]
    #Sets a filter on the KVFinder output, excluding cavities with smaller volumes than this parameter.
    volume_cutoff = 5.00
    #Defines the search radius for the ligand adjustment mode.
    ligand_cutoff = 5.00
    #Defines an removal distance when defining cavities by comparing Probe In and Probe Out surfaces. Default: 2.4 angstroms.
    removal_distance = 2.40

    [SETTINGS.visiblebox]
    #Coordinates of the box vertices that defines the search space. Only four points are necessary to define the box.
    bP1 = {bX1 = 0.00, bY1 = 0.00, bZ1 = 0.00}
    bP2 = {bX2 = 0.00, bY2 = 0.00, bZ2 = 0.00}
    bP3 = {bX3 = 0.00, bY3 = 0.00, bZ3 = 0.00}
    bP4 = {bX4 = 0.00, bY4 = 0.00, bZ4 = 0.00}

    [SETTINGS.internalbox]
    #Coordinates of the internal box vertices.
    P1 = {X1 = -4.00, Y1 = -4.00, Z1 = -4.00}
    P2 = {X2 = 4.00, Y2 = -4.00, Z2 = -4.00}
    P3 = {X3 = -4.00, Y3 = 4.00, Z3 = -4.00}
    P4 = {X4 = -4.00, Y4 = -4.00, Z4 = 4.00}

    [SETTINGS.eletrostatic]
    #Solve Poisson-Boltzmann equation with npbe, lpbe or lrpbe.
    pbe = "lpbe"
    #Type of boundary condition used to solve Poisson-Boltzmann equation. Options: zero, sdh, mdh.
    bcfl = "sdh"
    #Method by which the biomolecular point charges are mapped to the grid for a multigrid Poisson-Boltzmann equation. Options: spl0, spl2
    spl4.
    chgm = "spl2"
    #Model used to construct the dieletric and ion-accessibility coefficients. Options: mol, smol, spl2, spl4.
    srfm = "smol"
    #Dieletric constant of the solute molecule.
    pdie = 2.00
    #Dieletric constant of the solvent.
    sdie = 78.54
    #Number of quadrature points per square angstrom to use in calculation surface terms. Default: 10.0 angstroms.
    sdens = 10.00
    #Radius of the solvent molecules
    srad = 1.40
    #Size of the support for spline-based surface definitions. Default: 0.3 angstrom.
    swin = 0.30
    #Temperature for the calculation. Default: 298.15 K.
    temp = 298.15
```

Anexo B3: Fragmento dos arquivos de resultados <PDB>*.KVFinder.results.txt* do programa KVFinder.

```
KVFinder Input = /home/user/KVFinder/tutorial/5HKV.pdb
KVFinder Output = /home/user/KVFinder/tutorial/KV_Files/5HKV.KVFinder.output.pdb

#KVFinder Results:
Cavity KAA: Volume = 5.62 Angstrons^3
Cavity KAB: Volume = 26.35 Angstrons^3
Cavity KAA: Area =  10.72 Angstrons^2
Cavity KAB: Area =  34.16 Angstrons^2

#Interface Residues for Each Cavity:
Cavity KAA 26 D
Cavity KAA 235 C
Cavity KAA 234 C
Cavity KAA 233 C
Cavity KAA 232 C
Cavity KAA 97 C
Cavity KAA 96 C
Cavity KAA 95 C
Cavity KAA 50 C
Cavity KAB 293 D
Cavity KAB 292 D
Cavity KAB 291 D
Cavity KAB 290 D
Cavity KAB 40 D
Cavity KAB 25 D
Cavity KAB 24 D
Cavity KAB 23 D
Cavity KAB 237 C
Cavity KAB 236 C
Cavity KAB 235 C
Cavity KAB 50 C
Cavity KAB 49 C
```

Anexo B4: Fragmento do arquivos de resultados <PDB>*.KVFinder.results.toml* do programa parKVFinder.

```
#TOML results file for KVFinder software

title = "KVFinder results file"

[FILES_PATH]
INPUT = /home/user/KVFinder/input/1FMO.pdb
OUTPUT = /home/user/KVFinder/input/KV_Files/1FMO/1FMO.KVFinder.output.pdb
LIGAND =

[PARAMETERS]
RESOLUTION = Low
STEP SIZE = 0.60

[RESULTS]
#Volume, area, depth and interface residues information for each cavity

    [RESULTS.VOLUME]
    #Volume unit is cubic angstrom
    KAA = 151.85
    KAB = 42.77

    [RESULTS.AREA]
    #Area unit is square angstrom
    KAA = 121.32
    KAB = 50.89

    [RESULTS.DEPTH]
    #Maximum depth unit is angstrom
    KAA = 4.24
    KAB = 2.40

    [RESULTS.VOLUME_DEPTH]
    #Volume depth unit is angstrom
    KAA = 1014.70
    KAB = 163.37

    [RESULTS.HYDROPATHY]
    #Average hydropathy for each hydrophobicity scale provided to KVFinder
    scales_names = ["HessaHeijne","MoonFleming","KyteDoolitte","ZhaoLondon","WimleyWhite"]
    KAA = [0.30,-0.87,1.35,0.82,0.37]
    KAB = [0.30,0.21,-0.22,-0.09,0.01]

    [RESULTS.ELECTROSTATIC]
    #Average electrostatic potential
    KAA = 2.45
    KAB = 0.44

    [RESULTS.RESIDUES]
    #Interface residues for each cavity
    #[residue number,"chain identifier","residue name","residue class"]
    KAA = {[14,"E","S","R3"],[15,"E","V","R1"],[18,"E","F","R2"],[19,"E","L","R1"],[100,"E","F","R2"],[152,"E","L","R1"],
    [155,"E","E","R4"],[156,"E","Y","R2"],[292,"E","K","R5"],[302,"E","W","R2"],[303,"E","I","R1"],[306,"E","Y","R2"]}
    KAB = {[18,"E","F","R2"],[22,"E","A","R1"],[25,"E","D","R4"],[26,"E","F","R2"],[29,"E","K","R5"],[97,"E","A","R1"],
    [98,"E","V","R1"],[99,"E","N","R3"],[156,"E","Y","R2"]}

        [RESULTS.RESIDUES.DISTRIBUTION]
        #Residues type and class distribution
        #[["residue code",frequency],["residue class",frequency]]
        KAA = {[["A",0],["R",0],["N",0],["D",0],["C",0],["Q",0],["E",1],["G",0],["H",0],["I",1],["L",2],["K",1],["M",0],["F"
        2],["P",0],["S",1],["T",0],["W",1],["Y",2],["V",1],["X",0]],[["R1",4],["R2",5],["R3",1],["R4",1],["R5",1],["RX",0]]}
        KAB = {[["A",2],["R",0],["N",1],["D",1],["C",0],["Q",0],["E",0],["G",0],["H",0],["I",0],["L",0],["K",1],["M",0],["F"
        2],["P",0],["S",0],["T",0],["W",0],["Y",1],["V",1],["X",0]],[["R1",3],["R2",3],["R3",1],["R4",1],["R5",1],["RX",0]]}
```

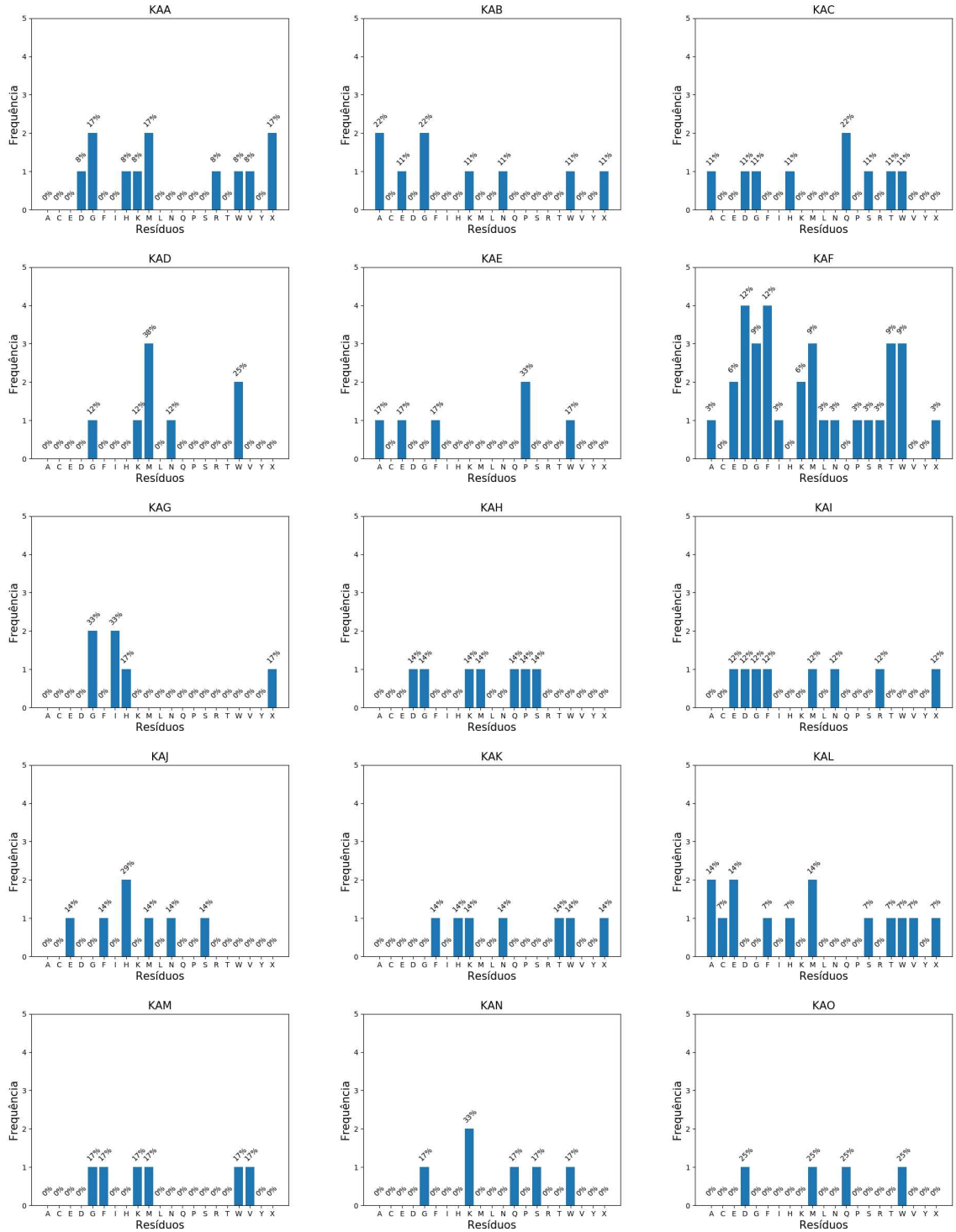
ANEXO C: DIRETIVAS DE COMPILAÇÃO DO OPENMP

Anexo C1: Definição das diretivas de compilação do OpenMP utilizadas no parKVFinder.

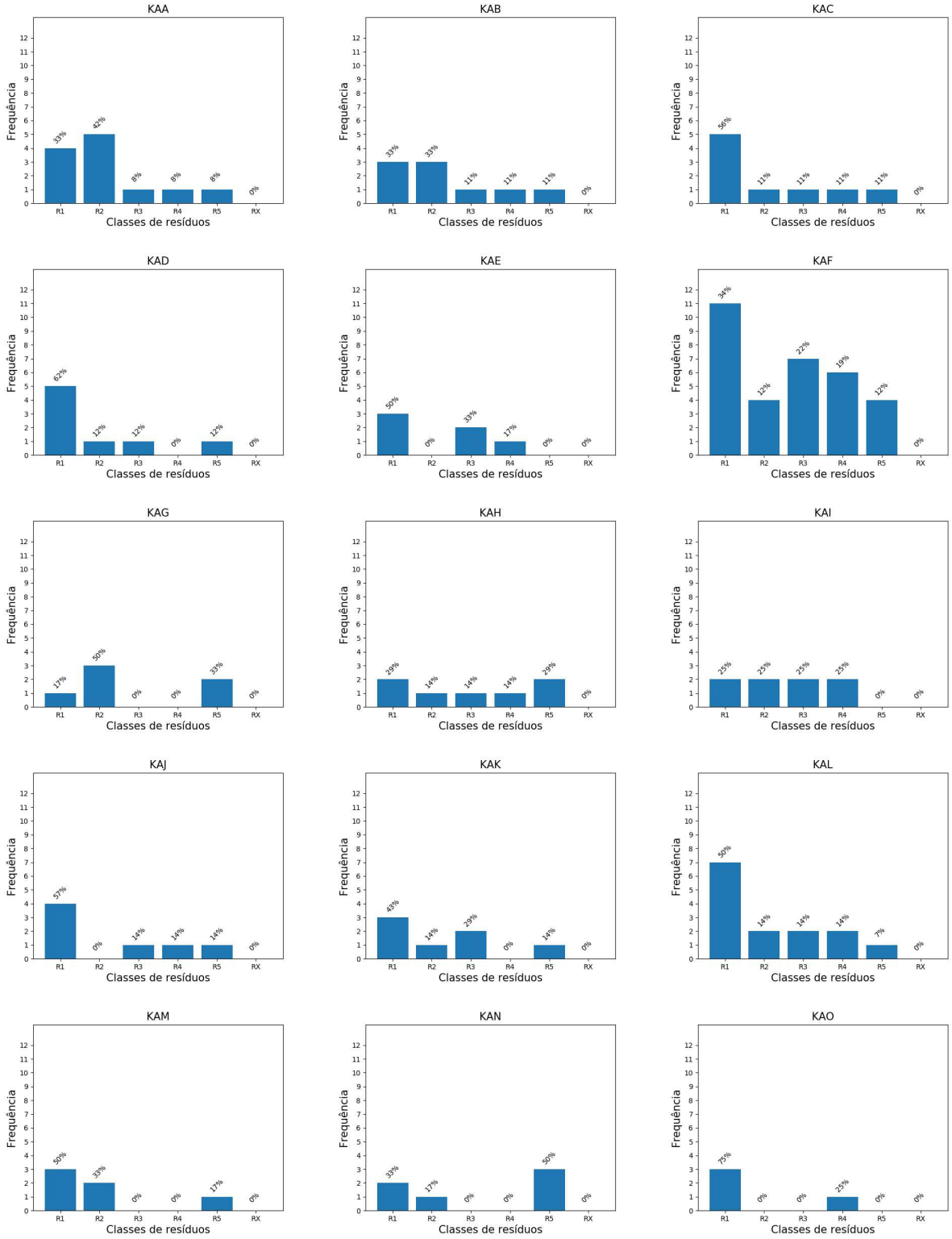
Diretiva de compilação	Definição
#pragma omp for colapse(n)	Transforma os seguintes n laços de repetição alinhados aninhados em apenas um laço de repetição linear
#pragma omp for schedule(dynamic)	Divide os laços de repetição são divididas em blocos e dinamicamente programados entre as tarefas; quando uma tarefa termina um bloco, outro é dinamicamente atribuído outro
#pragma omp for schedule(dynamic) nowait	Os blocos do laço de repetição atribuídos dinamicamente não são sincronizados ao final do laço de repetição

ANEXO D: HISTOGRAMAS CONSTITUCIONAIS DO PARKVFINDER

Anexo D1: Histogramas dos resíduos formadores das cavidades encontradas na estrutura da proteína quinase A.



Anexo D2: Histogramas das classes de resíduos formadores das cavidades encontradas na estrutura da proteína quinase A.



ANEXO E: DECLARAÇÕES

Anexo E1: Declaração de Bioética e/ou Biossegurança

COORDENADORIA DE PÓS-GRADUAÇÃO
INSTITUTO DE BIOLOGIA
Universidade Estadual de Campinas
Caixa Postal 6109. 13083-970, Campinas, SP, Brasil
Fone (19) 3521-6378. email: cpgib@unicamp.br

DECLARAÇÃO

Em observância ao **§5º do Artigo 1º da Informação CCPG-UNICAMP/001/15**, referente a Bioética e Biossegurança, declaro que o conteúdo de minha Dissertação de Mestrado, intitulada "***Prospecção e caracterização de cavidades supramoleculares***", desenvolvida no Programa de Pós-Graduação em Biociências e Tecnologia de Produtos Bioativos do Instituto de Biologia da Unicamp, não versa sobre pesquisa envolvendo seres humanos, animais ou temas afetos a Biossegurança.

Assinatura: João Victor da Silva Guerra
Nome do(a) aluno(a): João Victor da Silva Guerra

Assinatura:
Nome do(a) orientador(a): Paulo Sergio Lopes de Oliveira

Data: 24/10/2018

Anexo E2: Declaração de direitos autorais

Declaração

As cópias de artigos de minha autoria ou de minha co-autoria, já publicados ou submetidos para publicação em revistas científicas ou anais de congressos sujeitos a arbitragem, que constam da minha Dissertação/Tese de Mestrado/Doutorado, intitulada **Prospecção e caracterização de cavidades supramoleculares**, não infringem os dispositivos da Lei n.° 9.610/98, nem o direito autoral de qualquer editora.

Campinas, 01 de Agosto de 2019

Assinatura : João Victor da Silva Guerra
Nome do(a) autor(a): **João Victor da Silva Guerra**
RG n.° 49.716.592-2 SSP/SP

Assinatura :
Nome do(a) orientador(a): **Paulo Sérgio Lopes de Oliveira**
RG n.° 21.675.298-8 SSP/SP