Desenvolvimento de um algoritmo para identificação e
caracterização de cavidades em regiões específicas de estruturas
tridimensionais de proteínas.

Dissertação de Mestrado
Aluno: Saulo Henrique Pires de Oliveira
Orientador: Dr. Paulo Sergio Lopes de Oliveira
Número USP: 5483388

Programa de Pós-Graduação em Bioinformática da Universidade de São Paulo
Universidade de São Paulo, São Paulo
Abril de 2011

Dedico este trabalho aos meus pais, Renato Pires de Oliveira e Margarete Zapparoli de Oliveira, por todo o apoio e amor, indispensáveis ao sucesso de minha carreira profissional.

Resumo

A identificação e caracterização geométrica e físico-química de espaços vazios na estrutura tridimensional de proteínas é capaz de agregar informações importantes para guiar o desenho racional de drogas e a caracterização funcional de sítios de ligação e sítios catalíticos. Dessa forma, algumas ferramentas computacionais foram desenvolvidas nas últimas duas décadas, visando efetuar essas caracterizações. Contudo, as ferramentas existentes lidam com uma série de limitações, das quais merecem destaque a falta de precisão, falta de capacidade de integração em protocolos de larga escala, falta de capacidade de customização e a falta de uma caracterização eletrostática.

Tendo em mente estas limitações, desenvolvemos uma nova ferramenta, denominada KV-Finder, com o objetivo de estender as funcionalidades dos programas existentes, fornecendo assim uma caracterização sistemática mais eficiente e mais informativa dos espaços vazios da estrutura tridimensional de proteínas.

Através de uma modelagem matricial baseada em um direcionamento realizado pelo usuário, nossa ferramenta identifica e caracteriza espaços vazios em topologias proteicas. O utilitário é capaz de quantificar o volume, a forma, a extensão de sua superfície, os resíduos proteicos que interagem com os espaços vazios e um mapa de cargas parciais da superfície encontrada.

Nossa rotina foi integrada com ferramentas gráficas de modelagem molecular, fornecendo uma interação fácil e eficiente com o usuário. A validação de nosso algoritmo foi realizada em um conjunto de proteínas cujos diversos tipos de espaços vazios englobam os mais variados sítios de ligação e sítios catalíticos. O cálculo do volume de cavidades enzimáticas foi efetuado em larga escala, acompanhando a evolução do tamanho de bolsões na superfamília ALDH.

Com relação aos outros softwares existentes, nossa ferramenta apresenta uma série de vantagens das quais merecem destaque menor tempo de execução, maior precisão, maior acessibilidade e facilidade de integração com outros programas, além das características únicas de permitir que a busca ocorra em regiões específicas dentro da proteína e de realizar um mapeamento parcial de cargas da superfície encontrada.

Abstract

The identification and characterization of geometrical and physical-chemical properties in protein vacant spaces aggregates important information for steering rational drug designing and functional characterization of binding and catalytic sites. Therefore, several softwares have been developed during the past two decades in order to perform such characterization. Nevertheless, the existing tools still present a series of limitations such as lack of precision, lack of integrability in large scale protocols, lack of customization capacity and the lack of a proper electrostatic depiction.

We developed a new software, dubbed KV-Finder, in order to complement and extend the functionality of existing softwares, providing a systematic and more descriptive portrayal of protein vacant spaces. By employing a user-driven matrix modeling, our tool identifies and characterizes empty spaces in all sorts of protein topologies. The software quantifies the volume, the area and the shape of the surface, the residues that interact with the vacant spaces and a partial charge map of the computed surface.

Our routine was integrated with a graphical molecular modeling software, providing the user with a simple and easy-to-use interface. KV-Finder has been validated with a distinct set of proteins and binding sites. The volume computation was carried in large scale, accompanying the evolution of the pocket volume in the ALDH superfamily.

Compared with existing software, KV-Finder presents greater precision, greater accessibility and ease of integration in large scale protocols and visualization softwares. Also, the software possesses unique and innovative features such as the ability to segment and subsegment the empty spaces, an electrostatic depiction and a ligand interaction highlight feature.

Agradecimentos

Agradeço ao meu orientador, o Dr. Paulo Sergio Lopes de Oliveira, sem o qual a realização desta dissertação e do meu trabalho não seria possível.

Agradeço ao nosso colaborador, o Dr. Tiago Sobreira, pelas grandes contribuições ideológicas e por todo o apoio durante a execução do meu projeto.

Agradeço ao nosso colaborador, o Dr. José Xavier Neto, pelos conselhos e contribuições que ajudaram a enriquecer o projeto e o texto contido neste trabalho.

Agradeço ao meu colega, o Dr. Márcio Augusto Afonso de Almeida, por todo apoio, pelas grandes discussões e pela companhia durante a execução deste trabalho.

Agradeço aos meus colegas, o Dr. Pedro Edson Guimarães, o Ms. Júlio Cesar Nunes e o mestrando Rodrigo Vargas pela companhia ao longo do meu mestrado.

Agradeço a FAPESP pelo financiamento do meu projeto, sem o qual, certamente, o desenvolvimento do meu trabalho não seria possível.

Agradeço ao INCOR e ao LNBIO por fornecerem condições de trabalho ideais para a realização das minhas análises e implementação do meu projeto.

Agradeço a Sra. Patrícia Martorelli, por todo o apoio com relação aos trâmites burocráticos exigidos pela universidade de São Paulo com relação ao meu mestrado.

Agradeço ao Prof. Dr. Alan Mitchell Durham e a todos os membros da comissão de pós-graduação do programa inter-unidades de bioinformática da Universidade de São Paulo.

Agradeço a minha família e amigos por todo o apoio e carinho durante estes dois anos tão essenciais para o desenvolvimento da minha carreira.

Sumário

1. Introdução.....	9
1.1 Conteúdos desta Dissertação.....	9
1.2 Introdução de conceitos.....	10
1.3 Banco de Dados de Estruturas Tridimensionais de Proteínas.....	14
2. Estado da Arte.....	16
2.1 Matriz Tridimensional Booleana.....	16
2.2. Abordagens envolvendo Forma- α	20
2.3 Outras Abordagens.....	23
3. Justificativa.....	24
4. Objetivos.....	25
4.1 Objetivos específicos:.....	25
5. Métodos.....	27
5.1 A Interface Gráfica.....	29
5.2 Entrada de Dados.....	32
5.3 Carregamento dos Dicionários	34
5.3.1 Detalhes da Implementação.....	35
5.4 Extração das Coordenadas Atômicas	36
5.5 Montagem da Matriz Booleana.....	37
5.6 Preenchimento da Matriz Booleana.....	39
5.6.1 Modelos de Superfícies Proteicas.....	39
5.6.2 Primeira Etapa da rotina de Preenchimento.....	41
5.6.2.1 Detalhes da Implementação.....	42
5.6.3 Segunda Etapa da rotina de Preenchimento.....	47
5.6.3.1 Detalhes da Implementação.....	48
5.7 Ajuste ao Ligante.....	50
5.7.1 Detalhes da Implementação.....	51
5.8 Agrupamento de Espaços Vazios.....	51
5.8.1 Detalhes da Implementação.....	51
5.9 Cálculo do Volume.....	55
5.10 Cálculo da Superfície.....	55
5.10.1 Detalhes da Implementação.....	56
5.11 Mapeamento Eletrostático.....	57
5.12 Rotinas Adicionais.....	58
5.12.1 A Rotina de Busca a priori.....	58
5.12.1.1 Detalhes da Implementação.....	59
5.12.2 Identificação de resíduos expostos.....	60
5.12.2.1 Detalhes da Implementação.....	60
5.13 Criação de um modelo de saída dos dados obtidos.....	60
6. Resultados.....	62
6.1 Descrição espacial e eletrostática de uma fenda superficial.....	62
6.2 Particionamento de um sítio ativo e integração em análise de larga escala	68
6.3 Alto grau de detalhe na descrição de espaços vazios.....	73
6.3.1 Ajuste dos espaços vazios a um conjunto de ligantes.....	75
6.4 Particionamento de um Sítio de Ligação Complexo.....	77
6.5 Análise de Precisão e Tempo de Execução.....	82
7. Discussão.....	86
7.1 Estudo do Método.....	87
7.2 Tratamento de Elementos topográficos Superficiais.....	88

7.3 Caracterização Eletrostática.....	89
7.3 Particionamento de Elementos Topológicos.....	90
7.4 Maior Precisão no Cálculo de Volume.....	91
7.5 Integração com Softwares de Visualização.....	93
7.6 Considerações Finais.....	93
8. Conclusões.....	95
9. Bibliografia.....	96

1. Introdução

1.1 Conteúdos desta Dissertação

Nesta seção, descreveremos brevemente os conteúdos desta dissertação:

-Nos capítulo 1 e 2, introduzimos conceitos e discutimos o estado da arte na detecção e caracterização de espaços vazios proteicos.

-Nos capítulos 3 e 4, discutimos os objetivos e justificativas para a realização deste projeto.

-No capítulo 5, elaboramos uma descrição detalhada dos métodos empregados para a elaboração e implementação da nossa ferramenta computacional.

-No capítulo 6, apresentamos os resultados obtidos na análise estrutural de diversas proteínas. Neste capítulo, também apresentamos uma análise de precisão e eficiência do nosso software comparado a outras ferramentas existentes.

-Nos capítulos 7 e 8, os resultados são discutidos e as conclusões desta dissertação são construídas com base nos resultados encontrados e na comparação com as ferramentas existentes.

-No capítulo 9, apresentamos a nossa bibliografia.

1.2 Introdução de conceitos

Proteínas são macromoléculas essenciais para o funcionamento de qualquer célula existente. Estas macromoléculas estão envolvidas na grande maioria dos processos biológicos, desempenhando papéis cruciais para manutenção e funcionamento da maquinaria celular. Desta forma, o estudo de proteínas é extremamente importante do ponto de vista biológico.

Proteínas são formadas através da interação de moléculas menores, denominadas aminoácidos. Aminoácido é um nome que designa uma molécula que apresenta um grupo amina e um grupo carboxílico ligados a um mesmo carbono, denominado carbono alfa. A variedade de aminoácidos advém da cadeia lateral, que também se liga diretamente no carbono alfa.

O conjunto de aminoácidos canônicos contém vinte moléculas distintas. Contudo, os constituintes proteicos podem sofrer alterações após a tradução, tais como fosforilações, metilações e glicosilações. Portanto, considerando as modificações pós-traducionais, o conjunto de constituintes proteicos aproxima-se de trezentas moléculas diferentes.

Os aminoácidos de uma proteína interagem primariamente através de ligações covalentes denominadas ligações peptídicas: uma ligação entre o grupo amina de um aminoácido e o grupo carboxílico de outro aminoácido. As proteínas são compostas por sucessivas ligações peptídicas. Elas apresentam dois terminais, denominados N- e C-terminal, devido ao grupo amina e grupo carboxílico, respectivamente, livres em suas extremidades.

A maioria das ligações covalentes permite rotações entre os átomos que elas

conectam, sendo portanto flexíveis. Dessa forma, as ligações peptídicas permitiriam com que a proteína assumisse diversas formas tridimensionais diferentes, denominadas conformações. Contudo, a estrutura tridimensional de proteínas geralmente é restrita a poucas conformações. Essa rigidez é oriunda de ligações não covalentes entre os átomos proteicos, tais como pontes de hidrogênio, interações de Van der Waals e interações eletrostáticas.

Portanto, as propriedades físico-químicas de uma proteína resultam da interação entre as propriedades físico-químicas de seus constituintes. Em outras palavras, o conjunto de aminoácidos de uma proteína determina sua forma, carga, polaridade, etc... Considerando que o conjunto de constituintes varia de vinte até trezentos componentes diferentes e que proteínas podem ter milhares de aminoácidos em sua composição, estas macromoléculas apresentam um altíssimo grau de variabilidade em seu caráter físico-químico.

Dessa forma, proteínas apresentam alta complexidade estrutural. A estrutura tridimensional dessas macromoléculas é repleta de fendas, invaginações, bolsões, túneis, entre outros tipos de elementos topológicos. Essas formas topológicas apresentam uma grande variação em suas propriedades espaciais e físico-químicas, conferidas pelo caráter volumétrico e eletrostático dos aminoácidos que constituem cada região proteica.

Estes elementos topológicos podem ser classificadas de acordo com suas propriedades e funcionalidades específicas. Espaços vazios e superficiais podem ser classificados em quatro tipos distintos (Figura 1):

-Túneis (Figura 1A): Os túneis são espaços vazios que se estendem ao longo de um dos eixos da estrutura proteica. Os elementos topográficos que pertencem a esta categoria apresentam mais de um ponto de acesso (boca) ao meio externo da proteína

[1].

-Cavidades (Figura **1B**): as cavidades são espaços vazios compreendidos no interior da estrutura proteica [1], contendo no máximo uma boca. As cavidades podem ser subdivididas em três grupos: os vazios enclausurados, os bolsões e as invaginações. Um vazio enclausurado é um tipo de cavidade que não possui boca [1]. Os bolsões apresentam uma boca com área relativamente grande ao tamanho da cavidade. As invaginações, por outro lado, apresentam uma boca com área relativamente pequena ao tamanho da cavidade.

-Elementos superficiais (Figura **1C**): englobam fendas, fissuras, gretas, e outras depressões que se estendem pela superfície da estrutura sem apresentar grande profundidade.

-Vales Proteicos (Figura **1D**): Os vales proteicos são elementos topográficos profundos que se estendem ao longo de um dos eixos da proteína, sendo altamente expostos ao meio externo. A principal diferença entre os vales e o túneis é que os vales proteicos não apresentam uma boca clara, dado seu alto grau de abertura.

Figura 1: Representação das quatro categorias de elementos topográficos na estrutura tridimensional de proteínas: túneis (azul), cavidades (cinza), elementos superficiais (amarelo) e vales proteicos (verde).

A diversidade topológica permite com que as proteínas interajam com os mais diversos tipos de compostos de maneira altamente específica. A interação entre um sítio de ligação de uma dada proteína e seus ligantes específicos é fortemente dependente das propriedades volumétricas e físico-químicas de ambas as moléculas, sendo que é necessário um alto grau de complementariedade entre proteína e ligante para que uma interação possa ocorrer [2]. Portanto, as irregularidades topológicas geralmente atuam como sítios de ligação para os vários tipos de ligante envolvidos nas mais diversas atividades catalíticas [3].

Proteínas empregam mecanismos altamente complexos para permitir o acesso de

ligantes nos mais variados tipos de espaços onde se alocam sítios de ligação [4]. Na maioria dos casos, a identificação de sítios ativos pode ser reduzida a um problema de caracterização espacial (forma, volume, extensão) e eletrostática de espaços vazios compreendidos no interior ou na superfície de uma dada estrutura proteica. Seguindo o modelo de chave-fechadura, um espaço vazio que mantenha um alto grau de complementaridade entre suas propriedades espaciais com as mesmas propriedades de um possível ligante é um bom candidato para sítio ativo. Portanto, espaços vazios na estrutura proteica podem apresentar significado funcional.

Por conta disso, durante os processos de caracterização funcional de proteínas com papel biológico desconhecido e desenvolvimento racional de novos fármacos seria crucial agregar o maior número de informações estruturais e eletrostáticas da superfície e dos espaços negativos de uma proteína alvo. Essa informação pode então ser utilizada para prever e comparar ligantes, no caso da caracterização funcional, e para desenhar e modelar fármacos que interajam com a proteína alvo, apresentando grande relevância científica.

1.3 Banco de Dados de Estruturas Tridimensionais de Proteínas

As informações estruturais de uma proteína podem ser obtidas experimentalmente através de técnicas como cristalografia de raio X [5] e ressonância magnética nuclear [6]. Como resultado, a posição central de cada átomo no espaço pode ser armazenada em um sistema de coordenadas virtual dentro de um arquivo texto padronizado.

Dessa forma, o Banco de Dados Proteicos (*Protein Data Bank*) [7], ou PDB, foi criado para armazenar arquivos contendo as informações estruturais obtidas experimentalmente. Esses arquivos seguem os padrões e critérios estabelecidos pelo

banco e são curados e disponibilizados gratuitamente no domínio web do PDB.

O formato de arquivo PDB é um formato padrão aceito em praticamente qualquer software de modelagem e visualização de proteínas. Este arquivo agrega informações espaciais e sequencias de cada um dos átomos de uma dada proteína, separados por resíduo.

As informações contidas no PDB abriram margem para o desenvolvimento de ferramentas computacionais que realizam a identificação e caracterização das propriedades espaciais (volume, forma, área) e eletrostáticas de espaços vazios na estrutura tridimensional de proteínas.

2. Estado da Arte

Diversos métodos que realizam a caracterização geométrica de espaços vazios na estrutura tridimensional foram desenvolvidos durante as últimas duas décadas. As abordagens existentes podem ser divididas em três categorias principais:

- Abordagens utilizando Matrizes tridimensionais booleanas (também chamada de grade de pontos tridimensional).

- Abordagens utilizando o conceito geométrico de forma- α .

- Outras abordagens.

As duas primeiras abordagens são as mais utilizadas por terem se demonstrado mais eficientes e precisas na caracterização de espaços vazios. O foco principal das ferramentas atuais é identificar sítios ativos sem possuir nenhuma informação funcional prévia sobre a proteína. Apesar de algumas ferramentas utilizarem informações eletrostáticas para realizar a predição de sítios, nenhuma ferramenta existente agrega essas informações à superfície encontrada.

2.1 Matriz Tridimensional Booleana

A abordagem de matriz tridimensional booleana consiste em utilizar uma malha de pontos para modelar o espaço proteico (Figura 2). Cada um desses pontos recebe um valor booleano (*VERDADEIRO* ou *FALSO*) que indica se o ponto está contido dentro da proteína (*FALSO*) ou se ele representa um espaço vazio (*VERDADEIRO*) (Figura 2B).

Os algoritmos que empregam esta abordagem são, em geral, exaustivos. O princípio destes algoritmos é percorrer cada ponto da matriz e computar se aquele ponto pertence ao interior da proteína ou não. Uma definição mais formal deste conceito será

apresentada na seção 5.6. Cada ponto da grade é equidistante dos seus vizinhos ortogonais e essa distância é geralmente referida como o passo da matriz.

Figura 2: Representação em duas dimensões da modelagem utilizada na abordagem de Matriz Tridimensional Booleana. Em marrom está representada uma proteína que apresenta uma fenda superficial. Os pontos *VERDADEIROS* estão representados em verde, e os pontos *FALSOS* estão representados em vermelho.

A grande limitação encontrada por esta abordagem é discriminar entre espaços vazios com relevância biológica e o meio externo ao espaço proteico (Figura 3). Diversos métodos foram propostos para lidar com essa distinção, mas a grande maioria acaba por excluir os elementos superficiais durante este processo.

Figura 3: Representação em duas dimensões do processo de diferenciação entre espaço externo de uma proteína e seus elementos topográficos. Após a diferenciação, os pontos da grade que correspondem ao espaço externo são excluídos da representação.

Um dos primeiros softwares desenvolvidos para identificar espaços vazios em proteínas foi o VOIDOO [1]. VOIDOO emprega a abordagem de matriz tridimensional para modelar o espaço e buscar por possíveis cavidades. Este software consegue calcular o volume e a área da superfície de cavidades simples.

Para discriminar entre o espaço externo da proteína e uma possível cavidade, VOIDOO utiliza uma rotina chamada algoritmo *zapping* que consiste em aumentar os raios de Van der Waals dos átomos da proteína alvo até que as cavidades sejam efetivamente isoladas do meio exterior. Devido a essa rotina, VOIDOO não consegue encontrar bolsões e elementos topográficos mais complexos, uma vez que existe grande perda de informação na execução do algoritmo de *zapping*.

O software POCKET [8] utiliza uma metodologia similar àquela utilizada pelo VOIDOO. Essa ferramenta também faz uso de uma grade tridimensional para buscar por cavidades e túneis em uma proteína. Para resolver o problema de reconhecimento do espaço externo, ele fornece ao usuário a opção de definir uma distância para limitar esse

espaço exterior. Ainda assim, essa solução não é ótima, podendo haver grande perda de informação. Esse algoritmo despreza interações diagonais entre pontos da grade, portanto ocorre uma perda ainda maior de informação espacial. Isso faz com que o método seja dependente da orientação espacial original, e orientações diferentes de uma mesma proteína geram mapeamentos diferentes.

Um novo algoritmo matricial independente de orientação proteica foi então desenvolvido, o algoritmo LIGSITE [9]. Uma de suas principais implementações, o software Pocket Finder [10], utiliza uma grade de pontos otimizada que considera as interações diagonais entre os pontos, além de considerar interações verticais. Portanto, essa ferramenta resolve uma série de desvantagens associadas aos seus predecessores.

Contudo, falta ao Pocket Finder capacidade de customização: a ferramenta só é capaz de calcular a superfície de acesso ao solvente, o passo da matriz é restrito a valores maiores que 1.0 Å, os usuários não podem definir parâmetros de filtragem, dicionários internos e limites de busca. Portanto, Pocket Finder é incapaz de realizar a partição de espaços negativos ou de realizar buscas em regiões específicas. A ferramenta também não realiza o tratamento das propriedades eletrostáticas dos espaços vazios.

O software Q-Site Finder [13] estendeu as funcionalidades do Pocket Finder ao empregar cálculos de energia para prever sítios de ligação. Q-Site Finder efetua a ligação de sondas hidrofóbicas de CH₃ à superfície gerada pelo Pocket Finder. Agrupamentos de sondas são então classificados por energia de ligação mais ou menos favorável e uma probabilidade do grupo de sondas ser um sítio de ligação é calculada baseado na soma total de energias de ligação de cada grupo. Em geral, Q-Site Finder é um software preciso que descreve túneis e a maioria das cavidades com precisão.

Contudo, a usabilidade do programa é limitada pela sua abordagem web, que apresenta dificuldades durante a integração com protocolos de larga escala e

customização de parâmetros. Além disso, o Q-Site Finder não permite que você defina uma caixa de busca, sendo incapaz de realizar buscas em regiões específicas, de particionar espaços vazios e incapaz de descrever fendas superficiais.

Tanto o Pocket Finder como o Q-Site Finder geram mapas físico-químicos de espaços vazios, fornecendo uma quantidade significativa de informações sobre possíveis sítios de ligação. Entretanto, suas funcionalidades podem ser melhoradas, permitindo que os usuários consigam agregar um conjunto ainda maior de informações dos sítios de ligação, de maneira mais fácil e eficiente.

2.2. Abordagens envolvendo Forma- α

Algumas ferramentas para detecção de cavidades empregam um conceito de geometria computacional denominado forma- α [12]. A definição de forma- α é uma formalização da noção intuitiva da forma de um grupo de pontos. Diversas formas- α podem ser derivadas de um mesmo grupo de pontos, variando-se o grau de detalhe desejado através de um parâmetro α (Figura 4). Este parâmetro define o raio de uma bola aberta, sendo que para atingir completude $\alpha = 0$ representa um ponto e $\alpha = \infty$ representa um meio espaço.

Definição: Dado um conjunto S de pontos no espaço e um parâmetro α , uma forma- α é o conjunto de todos os vértices e triângulos formados por dois ou três pontos de S para os quais existe uma bola aberta $\mathbf{b}$ de raio α , tal qual estes pontos estejam contidos na casca de $\mathbf{b}$ e a intersecção entre $\mathbf{b}$ e o conjunto S seja nula.

Figura 4: Representação em duas dimensões para quatro formas- α diferentes para o mesmo conjunto de pontos.

A seguir, iremos explicar de uma maneira mais ilustrativa o conceito de forma- α para duas dimensões. Considere um conjunto de pontos e um valor positivo de α . Para todos os pares de pontos, você estabelece um segmento de reta entre estes pontos se existe uma circunferência de raio α tal que o seu par de pontos pertença a borda da circunferência e tal que nenhum outro ponto esteja contido na circunferência (Figura 5).

Figura 5: Representação em duas dimensões do conceito de forma- α para um conjunto de pontos. Define-se um segmento de reta (preto) entre dois pontos (amarelo), se houver uma circunferência de raio α (ciano) na qual os dois pontos pertencem a borda e nenhum outro ponto está contido.

As ferramentas que empregam forma- α [13,14] para mapear as propriedades geométricas de espaços vazios em proteínas utilizam uma aproximação. Estes programas consideram as coordenadas do centro de cada um dos átomos da proteína como o conjunto de pontos a ser utilizado no algoritmo. Isso é uma aproximação porque os raios atômicos não são considerados e, portanto, a superfície encontrada não corresponde com nenhum dos modelos de superfície proteica existentes.

Estas ferramentas são capazes de calcular uma solução para o problema de mapeamento, reconhecendo também as áreas de abertura (boca) das cavidades e túneis. Contudo, encontram dificuldades ao mapear as propriedades físico-químicas da superfície, uma vez que existe uma dificuldade em atribuir a informação eletrostática aos vértices e triângulos computados. Além disso, esse algoritmo não consegue restringir a região da busca ou compartimentalizar os espaços negativos, dado que todos os átomos da proteína precisam ser empregados para obter a forma- α da estrutura proteica. A

caracterização geométrica de fendas superficiais é altamente imprecisa, uma vez que os raios atômicos não são considerados.

2.3 Outras Abordagens

Outras abordagens para mapear espaços negativos na estrutura tridimensional de proteínas incluem softwares como o SURFNET [15] que utiliza um método que toma um par de átomos e desenha esferas crescentes que tangenciem as esferas geradas pelos raios de van der Waals dos átomos. Infelizmente, o utilitário não apresenta muita flexibilidade com parâmetros e ocorre grande perda de informação na caracterização de espaços vazios mais complexos. Existem ainda ferramentas que tratam de situações específicas, tendo como principal exemplo o utilitário HOLE [16], que busca somente por canais iônicos.

3. Justificativa

A identificação e a caracterização de propriedades físico-químicas de espaços vazios na estrutura tridimensional de proteínas fornecem informações extremamente importantes para guiar a caracterização funcional de novos sítios de ligação e o desenvolvimento racional de novas drogas. Tendo em vista a importância biológica desse problema, foram criadas ferramentas computacionais que realizam a identificação e caracterização de cavidades em proteínas.

Contudo, as ferramentas existentes para buscar e mapear espaços negativos em estruturas proteicas apresentam uma série de limitações como falta de precisão na caracterização de fendas superficiais, falta de caracterização eletrostática, dificuldade de integração com softwares de visualização, falta de customização de parâmetros e problemas de integração em protocolos de larga escala. Estas características são realmente indesejáveis em quaisquer aplicações. Além disso, nenhuma ferramenta existente permite o particionamento de espaços vazios e a delimitação do espaço de busca.

Motivados pela necessidade de melhorar as ferramentas existentes, nós desenvolvemos um novo e interativo software, altamente customizável e de fácil integração em protocolos de larga escala. Nossa ferramenta, chamada de KV-Finder, é capaz de identificar todos os tipos de espaços vazios, incluindo as fendas superficiais, calculando o volume e a área da superfície obtida, bem como um mapa de cargas parciais. Além disso, KV-Finder permite definir uma caixa que restringe o espaço de busca, efetivamente particionando os espaços vazios.

4. Objetivos

Elaborar, implementar e validar um algoritmo computacional eficiente para detecção e caracterização das propriedades físico-químicas de espaços vazios em regiões específicas de estruturas tridimensionais (3D) de proteínas. O algoritmo proposto será capaz de identificar os espaços vazios, agrupar os espaços encontrados, filtrar espaços sem relevância biológica, calcular os seus volume, calcular a área de suas superfícies e gerar um mapa de cargas parciais da superfície encontrada.

4.1 Objetivos específicos:

1- Implementar um método computacional para a identificação de espaços vazios à partir de um grupo de coordenadas atômicas, filtrando os espaços que não contém significado biológico. Para tal, faremos uso de uma abordagem de grade tridimensional regularmente espaçada. Como diferencial, empregaremos uma caixa definida pelo usuário que irá delimitar a matriz tridimensional, definindo regiões específicas para que a busca seja realizada e permitindo o particionamento dos espaços encontrados.

2- Agrupar os pontos da grade delimitando elementos topográficos diferentes. Calcular o volume e a área da superfície de cada um dos elementos encontrados através de um método de integração numérica. Indicar quais átomos e resíduos da proteína interagem com cada um dos espaços vazios encontrados.

3- Permitir que informações sobre um conjunto de ligantes seja utilizada para guiar o desenvolvimento de drogas, explicitando quais porções dos elementos encontrados não

são exploradas pelos fármacos existentes.

4- Determinar a forma e a extensão da superfície dos elemento topográfico encontrados. Integrar nossa ferramenta com um software gráfico, permitindo a visualização da superfície, bem como um alto grau de interação com os usuários. Agregar a esta interface gráfica um menu interativo, tornando as caracterizações altamente customizáveis pelos usuários.

5- Mapear as propriedades eletrostáticas das superfícies encontradas, permitindo assim uma melhor caracterização físico-química dos elementos topográficos proteicos.

6- Validar nosso algoritmo através de aplicações biológicas, como a descrição de diferentes tipos de cavidades, fendas superficiais e túneis, representando diversos sítios de ligação de enzimas e receptores.

7- Integrar nossa ferramenta em um protocolo de larga escala, permitindo uma análise sistemáticas de um elemento topográfico encontrado em uma família de proteínas.

5. Métodos

Implementamos uma ferramenta, denominada KV-Finder, utilizando a linguagem de programação ANSI C e a gama de bibliotecas *built-in* da própria linguagem. Os códigos foram compilados em ambiente UNIX e Windows e, em sua versão atual, o software apresenta compatibilidade com os sistemas operacionais Linux, Windows XP, Windows 7 e Mac OS X .

Para permitir uma maior interatividade com o usuário final, um *script* codificado em Tcl-Tk [17] também foi implementado, integrando assim o programa desenvolvido com outro software bastante utilizado em modelagem e visualização molecular, o *Visual Molecular Dynamics* (VMD) [18]. Para visualização de resultados e dos espaços negativos mapeados, bem como para gerar as figuras contidas neste documento, foram utilizados os softwares Pymol [19] e o já mencionado VMD.

A rotina empregada pelo nosso algoritmo será minuciosamente descrita a seguir e será apresentada uma explicação detalhada dos parâmetros contidos no arquivo de entrada, necessários para a utilização correta do nosso software.

Figura 6: Fluxograma da rotina empregada pelo nosso software para identificar, agrupar e mapear espaços vazios na estrutura tridimensional de proteínas.

5.1 A Interface Gráfica

A grande diferença da nossa ferramenta para os outros softwares que também empregam a abordagem de Matriz tridimensional booleana jaz no fato de que o KV-Finder permite que o usuário defina uma caixa no espaço, limitando assim a região de busca. Como esse tipo de definição é altamente visual, foi necessário integrar nosso utilitário com uma ferramenta de modelagem e visualização molecular.

O VMD é um software livre, bem estabelecido dentro da área de bioinformática estrutural e largamente utilizado pela comunidade acadêmica. O utilitário permite que módulos externos implementados em Tcl/Tk sejam adicionados ao programa. Desta forma, optamos por empregar este recurso do VMD para implementar uma interface gráfica para o KV-Finder.

Ao carregar o script da interface gráfica do KV-Finder dentro do VMD, o usuário clica na região da proteína onde ele deseja que a busca seja realizada e os limites da caixa ajustável se formam ao redor desta região (Figura 6). O usuário pode então aumentar ou diminuir os limites da caixa e até mesmo rotacioná-la de maneira intuitiva através dos botões na própria interface.

A maioria dos programas existentes apresenta dificuldades em diferenciar entre elementos topográficos e o espaço externo da proteína. Muitas vezes, os elementos superficiais não são detectados e caracterizados. Ao utilizar uma caixa definida pelo usuário, contraposto a uma caixa arbitrária, fixa e pré-definida, nossa ferramenta permite com que o pesquisador use sua intuição para realizar buscas por esses elementos, garantindo que o espaço de interesse seja propriamente detectado e caracterizado.

Figura 7: A interface gráfica do KV-Finder (destacada em verde claro) dentro do software de visualização molecular VMD. A caixa ajustável (amarelo) pode ser observada ao redor de uma região de busca específica na estrutura tridimensional de uma proteína ALDH2 humana (vermelho). Através das barras da interface gráfica, o usuário pode ajustar as dimensões da caixa e rotacionar a caixa para cobrir a região de interesse. Os arquivos de entrada e saída podem ser definidos e alterados diretamente na interface.

Além de definir as dimensões da caixa, o usuário pode acessar um menu de opções (Figura 8) e alterar os parâmetros das caracterizações de maneira rápida e intuitiva. Dado o baixo tempo de execução do KV-Finder (uma análise quantitativa do tempo de execução pode ser encontrada na seção 6.5), os usuários podem alterar os parâmetros e observar os resultados em poucos segundos. Dessa forma, a interface gráfica do KV-Fin-

der permite que o usuário gere sucessivos mapeamentos, fazendo ajustes a cada caracterização e obtendo assim o resultado desejado.

Figura 8: Menu de opções da interface gráfica do KV-Finder. Diversos parâmetros que podem ser ajustados pelo usuário em nossa interface não podem ser ajustados nas outras ferramentas existentes.

Apesar da interface gráfica ser extremamente útil para a definição dos limites da caixa e dos parâmetros utilizados pela nossa ferramenta, o funcionamento do KV-Finder é independente da interface gráfica. Todos os parâmetros são armazenados em um arquivo texto que é passado como entrada para o módulo central do programa, permitindo assim que o programa possa ser integrado em protocolos de larga escala de maneira eficiente. Caso o usuário decida por não utilizar a interface gráfica, basta editar o arquivo texto de configuração dos parâmetros e executar o software pelo terminal.

5.2 Entrada de Dados

Os dados de entrada do KV-Finder são divididos em três tipos de arquivos diferentes:

- 1- Um arquivo texto contendo os parâmetros.
- 2- Um arquivo PDB contendo informações sobre a estrutura da proteína alvo.
- 3- Os arquivos de dicionário.

O arquivo texto contendo as informações sobre os parâmetros é a base da entrada de dados. Este arquivo contém os nomes do arquivo PDB e dos arquivos de dicionário, bem como os diversos parâmetros importantes para a execução do KV-Finder. De forma geral, estes parâmetros englobam o passo da matriz, qual o tipo de superfície a ser calculado, o tamanho da sonda e parâmetros de filtragem (tais parâmetros serão discutidos com maiores detalhes nas seções seguintes). Neste arquivo também se encontram as coordenadas de quatro vértices da caixa que serão utilizados na montagem da matriz booleana.

O formato de arquivo PDB contém grande parte das informações sobre a estrutura tridimensional da proteína alvo (Figura 9). Cada átomo proteico é representado através de uma linha no arquivo e em cada uma destas linhas estão descritas as coordenadas (x,y,z) do centro de cada átomo da proteína.

CRYST1	73.080	78.440	80.490	90.400	90.400	90.400	P 1	1		
ATOM	1	N	GLU	E	13	-6.693	-15.642	-14.858	1.00100.00	N
ATOM	2	CA	GLU	E	13	-6.730	-14.620	-15.897	1.0084.38	C
ATOM	3	C	GLU	E	13	-7.490	-13.357	-15.508	1.00100.00	C
ATOM	4	O	GLU	E	13	-7.388	-12.338	-16.197	1.0096.04	O
ATOM	5	CB	GLU	E	13	-7.166	-15.114	-17.306	1.00100.00	C
ATOM	6	CG	GLU	E	13	-7.485	-16.623	-17.424	1.00100.00	C
ATOM	7	CD	GLU	E	13	-6.267	-17.503	-17.618	1.00100.00	C
ATOM	8	OE1	GLU	E	13	-5.439	-17.684	-16.719	1.00100.00	O
ATOM	9	OE2	GLU	E	13	-6.178	-18.047	-18.817	1.0066.45	O
ATOM	10	N	SER	E	14	-8.248	-13.420	-14.393	1.00100.00	N
ATOM	11	CA	SER	E	14	-8.988	-12.252	-13.929	1.00100.00	C
ATOM	12	C	SER	E	14	-7.964	-11.135	-13.736	1.00100.00	C

Figura 9: Um excerto de um arquivo PDB. As informações indicadas por setas coloridas são importantes para o funcionamento da nossa ferramenta. Estas entradas representam:

- Um identificador numérico indicando o número do átomo na proteína (seta preta).
- Um identificador sequencial dos átomos dentro do resíduo (seta azul).
- Um identificador indicando a qual resíduo proteico este átomo pertence (seta verde).
- Um identificador indicando a qual cadeia da proteína o átomo pertence (seta vermelha) .
- Um identificador indicando o número do resíduo na proteína (seta cinza).
- As três coordenadas atômicas (setas e área magenta).
- O símbolo atômico (seta marrom).

Apesar de agregar muitas informações espaciais sobre a estrutura de uma proteína, duas informações cruciais para a caracterização físico-química de espaços vazios protei-

cos estão ausentes no arquivo PDB: o raio atômico e a carga parcial de cada um dos átomos presentes na proteína. Estes dados são fornecidos para o KV-Finder através dos arquivos de dicionários que serão descritos na próxima seção.

5.3 Carregamento dos Dicionários

O KV-Finder emprega três dicionários para detectar e efetivamente mapear vazios proteicos. O primeiro dos três dicionários empregados pelo nosso programa é uma lista que contém todos os raios de Van der Waals identificados pelo símbolo atômico do elemento correspondente (Figura **10C**). Esta informação é necessária para mapear as propriedades geométricas dos contornos e interiores de qualquer proteína e é complementar às informações do arquivo PDB.

Os outros dois dicionários estão associados entre si e estão ligados à carga parcial dos átomos representados pelo arquivo PDB. O arquivo do dicionário de resíduos (ADR) lista cada um dos resíduos que estão contidos na proteína alvo (Figura **10B**). O ADR padrão contém uma lista dos resíduos comumente encontrados na maioria das proteínas. Todavia, resíduos não convencionais podem ser facilmente adicionados apenas editando o arquivo texto que representa o dicionário.

O arquivo do dicionário de cargas (ADC) contém a carga parcial de todos os átomos da macromolécula alvo, separados pelos resíduos listados no ADR (Figura **10A**). O ADC padrão é derivado de um campo de força específico, Amber 9 [20]. Esse dicionário é utilizado posteriormente para mapear as cargas na superfície dos espaços vazios proteicos.

A

>ALA	
N	-0.4157
HN	0.2719
H	0.2719
CA	0.0337
HA	0.0823
H1	0.0823
CB	-0.1825
HB1	0.0603
HB2	0.0603
HB3	0.0603
C	0.5973
O	-0.5679

C

AC	2.00
AG	1.72
AL	2.00
AM	2.00
AR	1.88
AS	1.85
AT	2.00
AU	1.66
B	2.00
BA	2.00
BE	2.00
BH	2.00
BI	2.00
BK	2.00
BR	1.85
C	1.66
CA	2.00

B

ALA	ARG	ASN	ASP	ASH	CYS	CYM	CYX	GLN	GLU
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

Figura 10: Fragmentos de cada um dos arquivos de dicionário utilizados pela nossa ferramenta, KV-Finder. O dicionário de cargas (**A**) apresenta a carga parcial de cada átomo separado por resíduos. O dicionário de resíduos (**B**) lista os resíduos contidos no dicionário de cargas. O dicionário de raios de Van der waals (**C**) apresenta o raio atômico de todos os elementos atômicos.

5.3.1 Detalhes da Implementação

A primeira etapa do algoritmo empregado pelo KV-Finder visa extrair os dados contidos nos arquivos de dicionário, armazenando-os em duas tabelas internas. A primeira tabela, a tabela do dicionário de raios, foi implementada através de uma tabela de *hash* com espalhamento ótimo, permitindo assim uma recuperação rápida de informações. As cha-

ves desta tabela são os símbolos atômicos e os valores são os próprios raios de Van der Waals.

A segunda tabela é implementada através de um vetor de ponteiros que apontam para uma lista ligada. Cada posição do vetor corresponde a um resíduo contido no dicionário de resíduos. Os elementos da lista ligada são identificados pelo símbolo sequencial do átomo dentro do resíduo. Dados o resíduo e este identificador do átomo dentro do resíduo, a carga parcial de qualquer átomo pode ser recuperada através de uma varredura da tabela. A complexidade da busca de cargas é, portanto, da ordem do número de resíduos fornecidos no ADR, supondo o número de átomos por resíduo aproximadamente constante.

5.4 Extração das Coordenadas Atômicas

Depois de carregar os dicionários, a próxima etapa do nosso algoritmo é a extração das coordenadas dos átomos do arquivo PDB fornecido como entrada. Esta extração nada mais é do que uma leitura e armazenamento das informações atômicas contidas no arquivo PDB.

Durante a leitura, realizamos uma filtragem dos átomos que não interagem com a caixa definida pelo usuário (figura 11). Nossa modelagem trata cada átomo como uma esfera centrada nas coordenadas fornecidas pelo arquivo PDB. Como os limites da caixa são fornecidos na entrada, basta avaliar se as esferas atômicas interseccionam o espaço delimitado pelo usuário. Dado que o tempo de execução de diversas etapas do algoritmo empregado pelo KV-Finder é diretamente proporcional ao número de átomos fornecidos como entrada, esta filtragem acarreta em um ganho de eficiência.

Se uma esfera representando um átomo proteico interage com o espaço delimitado pela caixa, suas coordenadas centrais são armazenadas em um vetor de átomos **V**. Os símbolos contidos no PDB são empregados para realizar uma consulta nas tabelas dos dicionários, permitindo que a carga parcial e o raio de Van der Waals de cada átomo a ser inserido também sejam obtidas e armazenadas no vetor de átomos. A complexidade desta rotina é da ordem do número de átomos contidos no arquivo PDB.

Figura 11: Representação esquemática de alguns dos átomos filtrados (vermelho) e não filtrados (verde) pelo etapa de extração de coordenadas atômicas. As esferas são centradas nas coordenadas de cada átomo fornecidas pelo arquivo PDB. O raio de cada esfera é o raio fornecido pelo dicionário de raios. Em **A**, uma representação tridimensional, em que o espaço vazio é representada em azul. Em **B**, uma representação bidimensional. Os átomos representados em verde são inseridos no vetor de átomos **V**.

5.5 Montagem da Matriz Booleana

Dadas as quatro coordenadas da caixa fornecidas durante a entrada de dados, nosso algoritmo executa uma transformação no sistema de coordenadas para representar o espaço através de uma matriz tridimensional. O primeiro vértice fornecido, após uma

translação, irá compor a origem do sistema de coordenadas transformado. Os outros vértices representarão vetores que apontam na direção dos eixos do novo sistema.

A matriz tridimensional booleana **A** é definida como um conjunto de pontos do espaço. Cada posição da matriz representa um único ponto no espaço, confinado dentro dos limites da caixa e distando de um passo **h**, que é fornecido como parâmetro de entrada, dos seus seis vizinhos imediatos na matriz. O valor de cada entrada na matriz indica se o espaço representado por ela está vazio ou não.

Vamos então definir os quatro pontos fornecidos na entrada como **P₁**, **P₂**, **P₃** e **P₄**. Em seguida, iremos obter as dimensões da matriz, definidas abaixo:

$$\mathbf{m} = (\mathbf{x}_2 - \mathbf{x}_1) / \mathbf{h} \qquad \mathbf{n} = (\mathbf{y}_3 - \mathbf{y}_1) / \mathbf{h} \qquad \mathbf{o} = (\mathbf{z}_4 - \mathbf{z}_1) / \mathbf{h} ,$$

onde **x_i**, **y_i** e **z_i** são as coordenadas de **P_i**, para **i** em {1, 2, 3, 4}.

Continuando, iremos obter os ângulos de rotação entre a caixa e seus eixos, necessários para realizar a transformação de coordenadas empregada em nosso algoritmo. Estes ângulos são obtidos através das seguintes relações:

$$\theta = \arccos ((\mathbf{x}_2 - \mathbf{x}_1) / \|\mathbf{P}_2 - \mathbf{P}_1\|) \qquad \sigma = \arccos ((\mathbf{y}_3 - \mathbf{y}_1) / \|\mathbf{P}_3 - \mathbf{P}_1\|)$$

Cada entrada da matriz irá representar a posição fornecida por uma translação, uma rotação e uma contração do sistema original de coordenadas, definindo então **P₁**

como sua origem. Portanto, os pontos no sistemas de coordenadas original da proteína podem ser obtidos através de computações simples utilizando os índices da matriz. Esse método fornece uma grade de pontos compreendida dentro dos limites da caixa. Cada posição da matriz **A** é inicializada como um espaço vazio (valor *VERDADEIRO*).

Dada a entrada da matriz **A_{ijk}**, em que $0 \leq i < m$, $0 \leq j < n$ e $0 \leq k < o$, o ponto **P** = (**x_P**, **y_P**, **z_P**) do espaço que é representado por esta entrada pode ser calculado como segue:

$$x_P = ((i * h) * \cos \theta + (j * h) * \sin \sigma * \sin \theta - (k * h) * \cos \sigma * \sin \theta) + x_1$$

$$y_P = ((j * h) * \cos \sigma + (k * h) * \sin \sigma) + y_1$$

$$z_P = ((i * h) * \sin \theta - (j * h) * \sin \sigma * \cos \theta + (k * h) * \cos \sigma * \cos \theta) + z_1$$

5.6 Preenchimento da Matriz Booleana

5.6.1 Modelos de Superfícies Proteicas

Antes de detalhar as etapas da rotina de preenchimento da matriz booleana, faz-se necessária a discussão e definição de modelos de superfícies proteica. Diversos modelos foram desenvolvidos de modo a definir a superfície de proteínas e outras macromoléculas com base em sua estrutura tridimensional (figura 12).

O primeiro e mais simplório modelo para representar a superfície de proteínas foi denominado superfície de Van der waals. Este modelo considera a superfície proteica que

é resultante dos raios de Van der Waals dos átomos que constituem a proteína. Esta representação apresenta uma superfície repleta de pequenos espaços vazios, uma vez que os raios de Van der waals não preenchem totalmente o espaço proteico. Esse fenômeno é conhecido como efeito *canning* [1].

Figura 12: Representação dos diferentes modelos de superfície proteica. A superfície de Van der waals (vermelho) é obtida considerando o raio de Van der waals dos átomos da proteína. As superfícies de acesso ao solvente (branco) e superfície molecular (azul) são obtidas rolando uma sonda esférica sobre a superfície de Van der waals. A SAS é a superfície acessível ao centro da sonda. A Superfície molecular é a superfície tocada e delimitada pela borda da sonda.

Outros dois modelos amplamente difundidos são o modelo de superfície de acesso ao solvente (SAS) [21,22] e o modelo de superfície molecular (também chamada de superfície solvente-excludente ou SSE) [22]. As duas representações são relativamente similares e são modelos bem representativos da superfície proteica. Ambas abordagens envolvem a rolagem de uma sonda esférica sobre a superfície de Van de Waals. Essa esfera é chamada de sonda molecular e o raio utilizado para a sonda é, geralmente, o raio aproximado de uma molécula de água (1.4 Å). O raio da sonda e a superfície proteica a ser computada são parâmetros que podem ser alterados pelo usuário na interface gráfica do KV-Finder.

5.6.2 Primeira Etapa da rotina de Preenchimento

O próximo passo do nosso protocolo de execução consiste em preencher os espaços não vazios com todos os átomos contidos no vetor de átomos **V**. Para isso, a rotina de preenchimento é dividida em duas etapas. A primeira etapa está associada com o cálculo da superfície de acesso ao solvente. A segunda etapa tem como objetivo obter a superfície molecular a partir da SAS obtida na primeira etapa. Nesta seção, iremos descrever a primeira etapa da rotina de preenchimento da matriz booleana.

Como mencionado anteriormente, cada posição da matriz tridimensional contém um booleano que indica se o ponto correspondente no espaço é vazio ou não. Logo, uma definição precisa de espaço preenchido se faz necessária.

Seja $\mathbf{P}_{ijk}$ um ponto representado pela entrada $\mathbf{A}_{ijk}$ da matriz espacial, para $0 \leq i < m$, $0 \leq j < n$ e $0 \leq k < o$. O ponto $\mathbf{P}_{ijk}$ é considerado um espaço vazio se, para cada átomo **a** em **V**, uma esfera com raio de Van der Waals r_a , centrada nas coordenadas $\mathbf{P}_a$ de **a**, não intercepta uma esfera de raio da sonda **p** centrada em $\mathbf{P}_{ijk}$ (Figura 13). O que

é equivalente à seguinte fórmula:

$$A_{ijk} = \text{VERDADEIRO} \Leftrightarrow \text{Para todo } \mathbf{a} \text{ em } \mathbf{V}, \|\mathbf{P}_a - \mathbf{P}_{ijk}\| > (r_a + p)$$

Figura 13: A intersecção entre duas esferas não ocorre quando a soma de seus raios é menor do que a distância entre os seus centros.

5.6.2.1 Detalhes da Implementação

Para realizar a primeira etapa da rotina, nosso algoritmo percorre todos os átomos do vetor $\mathbf{V}$. Para cada átomo, é percorrida uma submatriz que circunscreve a esfera obtida pelo raio de van der Waals mais o raio da sonda centrada nesse átomo, atribuindo valores *FALSO* aos pontos não vazios da matriz (Figura 14).

Figura 14: A primeira etapa da rotina de preenchimento da matriz percorre o vetor de átomos da proteína. Dado um átomo do vetor, definimos uma submatriz ao redor deste átomo:

Em seguida, a rotina percorre todos os pontos dessa submatriz. Em cada um dos pontos, centramos uma esfera de sonda e verificamos se ocorre intersecção com a esfera que representa o átomo, como ilustrado nos exemplos a seguir:

Neste exemplo, não ocorre intersecção. Portanto, atribui-se *VERDADEIRO* ao ponto da matriz.

Ocorre Intersecção

Neste exemplo, ocorre intersecção. Portanto, atribui-se *FALSO* ao ponto da matriz.

FALSO

Após percorrer todos os pontos da submatriz, o espaço preenchido pelo átomo é representado como pontos de valor *FALSO* na matriz booleana:

Após repetir o processo anterior para todos os átomos contidos no vetor de átomos, obtemos uma matriz que representa os espaços vazios e não-vazios compreendidos dentro dos limites da caixa:

Observe que para cada átomo, o tempo de execução é proporcional somente ao número de pontos da submatriz, obtido pela expressão k/h^3 , onde h é o passo da matriz e k é uma constante maior que 0. Portanto, a complexidade dessa etapa é $O(N / h^3)$, onde N é o número de átomos em V .

5.6.3 Segunda Etapa da rotina de Preenchimento

Após a primeira etapa de preenchimento, obtemos como resultante a superfície de acesso ao solvente. Para obter a superfície molecular, definimos dois novos conceitos: pontos de fronteira e esferas limítrofes.

Um ponto de fronteira é representado por uma entrada de valor *VERDADEIRO* da matriz que apresenta pelo menos um ponto adjacente (Figura 15) de valor *FALSO*. Um ponto de fronteira é definido como segue:

Definição: Para $1 \leq a < m-1$, $1 \leq b < n-1$ e $1 \leq c < o-1$, um ponto Pf_{abc} é considerado um ponto de fronteira se existe A_{ijk} , tal que $A_{abc} = VERDADEIRO$ e $A_{ijk} = FALSO$, onde $a-1 \leq i \leq a+1$, $b-1 \leq j \leq b+1$ e $c-1 \leq k \leq c+1$.

Figura 15: Representação esquemática de pontos vizinhos e pontos adjacentes. O ponto representado em preto é considerado um ponto de fronteira se ele apresenta valor *VERDADEIRO* e se pelo menos um de seus pontos adjacentes apresenta valor *FALSO*.

Para cada ponto de fronteira, definimos uma esfera limítrofe como uma esfera de raio de sonda **p** centrada nas coordenadas representadas pelo ponto de fronteira. Estas esferas de fronteira representam a sonda que é rolada sobre a superfície de van der Waals. Para obter a superfície molecular, é necessário atribuir *VERDADEIRO* aos pontos da grade compreendidos dentro das esferas limítrofes.

5.6.3.1 Detalhes da Implementação

A segunda etapa da rotina de preenchimento itera sobre cada entrada da matriz booleana procurando e identificando os pontos de fronteira. Uma vez que um ponto de fronteira é encontrado, a rotina atribui *VERDADEIRO* a todas as entradas que representam pontos contidos na esfera limítrofe de raio **p**, centrada no ponto de fronteira

encontrado (Figura 16).

Como a rotina itera por todos os pontos da matriz e para cada ponto de fronteira a rotina executa $O(1/h^3)$ operações, a complexidade da segunda etapa da rotina de preenchimento é $O(m*n*o / h^3)$.

Figura 16: Representação esquemática em duas dimensões da matriz espacial. Os átomos são representados por esferas azuis e amarelas em que a porção amarela corresponde ao raio da sonda. Em **A**, temos os pontos considerados como vazios (verde) pela superfície de acesso ao solvente. Em **B**, os pontos roxos indicam os pontos de fronteira encontrados pela segunda etapa da rotina de preenchimento. Em **C**, os pontos contidos na esfera limítrofe (linha laranja) são reconsiderados como *VERDADEIRO*. Em

D, ilustramos a superfície molecular resultante do algoritmo de preenchimento.

5.7 Ajuste ao Ligante

Se algum ligante da proteína alvo é conhecido, pode ser desejável utilizar essa informação para refinar o mapeamento dos espaços vazios. Seria extremamente interessante poder visualizar quais regiões de um elemento topográfico estão próximas de um conjunto de ligantes fornecidos como entrada, uma vez que isso permite a identificação de quais regiões do elemento topográfico estão interagindo com o conjunto de ligantes e quais regiões permanecem inexploradas.

Nosso algoritmo apresenta uma rotina de ajuste do espaço mapeado à uma molécula, supondo que as coordenadas desse ligante sejam fornecidas na entrada como um arquivo PDB. O algoritmo de ajuste ajusta o valor de cada entrada na matriz espacial conforme descrito abaixo:

Seja $\mathbf{P}_{ijk}$ o ponto representado por uma entrada $\mathbf{A}_{ijk}$ da matriz espacial, para $0 \leq i < m$, $0 \leq j < n$ e $0 \leq k < o$. O ponto $\mathbf{P}_{ijk}$ é reavaliado como um espaço vazio se ele distar de todos os átomos $\mathbf{a}$ contidos no vetor de átomos do ligante $\mathbf{VL}$ mais do que uma distância de corte $\mathbf{D}$, provida pela entrada. Isso pode ser resumido nas relações:

$\mathbf{A}_{ijk}$ recebe *FALSO* $\Leftrightarrow$ Para todo $\mathbf{a}$ em $\mathbf{VL}$, $\| \mathbf{P}_a - \mathbf{P}_{ijk} \| > \mathbf{D}$, para $\mathbf{D} > 0$

ou

$\mathbf{A}_{ijk}$ recebe *FALSO* $\Leftrightarrow$ Para todo $\mathbf{a}$ em $\mathbf{VL}$, $\| \mathbf{P}_a - \mathbf{P}_{ijk} \| < -\mathbf{D}$, para $\mathbf{D} < 0$

5.7.1 Detalhes da Implementação

Nossa rotina percorre todas as entradas de **A**, verificando quais valores devem ser ajustados. Para realizar essa verificação é necessário percorrer o vetor de átomos do ligante. Portanto, a complexidade da rotina de ajuste é $O(m*n*o*NL)$, onde **NL** é o número de átomos em **VL**.

5.8 Agrupamento de Espaços Vazios

Após a atribuição correta dos valores de cada entrada, obtemos uma matriz que contém uma quantidade significativa de informação sobre o espaço compreendido dentro dos limites da caixa (Figura **18A**). Contudo, essa informação ainda precisa ser processada antes que possa gerar resultados. O próximo passo da nossa rotina agrupa os espaços vazios, filtrando pequenos buracos sem significância biológica do espaço negativo da proteína alvo.

5.8.1 Detalhes da Implementação

A rotina de agrupamento modela o espaço vazio como um grafo, permitindo assim uma solução eficiente para o problema. Cada entrada de valor *VERDADEIRO* da matriz é tratada como um vértice, cujas arestas são estabelecidas com seus 26 pontos adjacentes. Todavia, uma aresta apenas pode ser estabelecida se o ponto adjacente também for uma

entrada de valor *VERDADEIRO*. Dessa forma, a matriz **A** passa a ser representada por um grafo **G** durante a extração (Figura 17).

Figura 17: Representação esquemática de uma matriz espacial **A** e sua modelagem representada pelo grafo **G**.

O algoritmo de agrupamento realiza uma busca em profundidade [23,24] para identificar corretamente cada um dos espaços vazios. Durante a busca, cada vértice é etiquetado com um inteiro correspondente ao número de identificação do espaço vazio. Após a busca, cada espaço vazio estará identificada como um componente conexo do grafo (Figura 18B).

Para realizar o agrupamento, nosso algoritmo itera por todas as posições da matriz

A, realizando a busca em profundidade para cada entrada vazia ainda não etiquetada. Cada ponto visitado é contado e a contagem total é armazenada, resultando em um vetor que contém o número total de pontos para cada espaço negativo. Cada entrada só é visitada apenas uma vez durante o algoritmo. Portanto, a complexidade dessa rotina é $O(m*n*o)$.

Um parâmetro de filtragem **f** precisa ser fornecido como entrada, consequentemente determinando o volume mínimo necessário para considerar um grupo de pontos como um espaço vazio com possível significado biológico. Qualquer grupo de pontos cujo volume é menor do que **f** é filtrado da matriz espacial realizando uma nova busca em profundidade que altera o valor de cada um dos pontos do grupo para *FALSO* (Figura 18C).

A

Figura 18A: O resultado da rotina de preenchimento é um grupo de pontos no espaço. A rotina de agrupamento define grupos de pontos como elementos topográficos distintos.

B

Figura 18B: Representação esquemática dos elementos topológicos distintos encontrados pela rotina de agrupamento. Entretanto, seria desejável filtrar espaços topológicos muito pequenos, possivelmente sem relevância biológica.

C

Figura 18C: Representação esquemática do resultado da rotina de agrupamento após o processo de filtragem.

5.9 Cálculo do Volume

A próxima etapa do nosso algoritmo estima o volume de cada um dos elementos topológicos identificados pela rotina de agrupamento. O KV-Finder utiliza um método de aproximação numérica para estimar o volume dos espaços vazios encontrados. Cada ponto da nossa matriz é tratado como um voxel, ou seja, um cubo com aresta h e volume h^3 . Como a busca em profundidade da rotina anterior já calculou o número de pontos em cada grupo, podemos estimar o volume apenas multiplicando o número de pontos pelo volume do voxel, o que acontece sem custo computacional adicional.

A área da superfície será calculada de maneira análoga, após a obtenção da superfície na rotina que será descrita na seção seguinte. O cálculo da área da superfície considera quadrados de lado h e área h^2 e a área pode ser estimada multiplicando o número de pontos encontrado pela área do quadrado, sem custo computacional adicional.

5.10 Cálculo da Superfície

A representação espacial resultante da rotina anterior nem sempre é a mais adequada para descrever os espaços vazios, uma vez que ela depende muito do modo em que a caixa foi definida. Uma caixa mal desenhada resultaria em um grande espaço cúbico vazio anexado ao espaço vazio (Figura **19B**). Graficamente, esse anexo é indesejável e dificulta qualquer análise visual dos resultados. Dessa forma, o modelo gráfico existente foi aperfeiçoado de modo a tratar essa desvantagem.

O próximo passo da nossa rotina define uma nova representação do espaço, considerando apenas a superfície dos vazios proteicos. Este modelo evita com que um grande número de pontos formem um cubo anexo a um agrupamento de pontos vazios

encontrados, ou seja, evita engrandecimentos indesejáveis provenientes de uma caixa mal alocada e melhora a resolução de bolsões e fendas, gerando uma representação gráfica mais adequada. Para isso, nosso algoritmo emprega uma transformação da matriz espacial, considerando apenas os pontos de fronteira como *VERDADEIROS*.

A

B

Figura 19: A diferença entre a superfície tratada (**A**) e a superfície não tratada (**B**) de um bolsão na superfície de uma proteína quinase cAMP-dependente de camundongo (Código PDB: 1FMO).

5.10.1 Detalhes da Implementação

A rotina de transformação da superfície itera por todas as posições de **A**, etiquetando as entradas que representam pontos de fronteira. Depois, através de uma segunda iteração, essa rotina atribui a cada entrada não etiquetada um valor *FALSO*, obtendo assim apenas os pontos que representam a superfície do espaço negativo. A complexidade da rotina de transformação da superfície é $O(m*n*o)$.

5.11 Mapeamento Eletrostático

O próximo passo do nosso algoritmo consiste em mapear as propriedades eletrostáticas da superfície encontrada. Esta informação é útil para compreender os mecanismos de interação entre sítio de ligação e ligantes, para o desenho racional de drogas e para a identificação de novos sítios ativos.

Para realizar o mapeamento parcial das cargas, define-se uma matriz *double* **M**, cujas dimensões são **m** x **n** x **o**. Cada entrada de **M** contém o valor da carga **C_a** do átomo **a** de **V**, em que essa posição representa um ponto contido na esfera pela qual o átomo **a** é representado. Se a entrada representa um espaço vazio, essa posição da matriz recebe o valor *NULL*.

Seja **P_{ijk}** o ponto representado por uma entrada **M_{ijk}** da matriz de cargas, para $0 \leq i < m$, $0 \leq j < n$ e $0 \leq k < o$. Nosso algoritmo atribui à entrada **M_{ijk}** os seguintes valores iniciais:

M_{ijk} recebe **C_a** se existe **a** em **V**, em que $\| \mathbf{P}_a - \mathbf{P}_{ijk} \| < r_a$

<

M_{ijk} recebe *NULL* caso contrário.

Após a inicialização da matriz **M**, é necessário obter o valor das cargas para as posições que correspondem a pontos de fronteira das cavidades na matriz **A**. Dessa forma, itera-se por todas as entradas da matriz espacial, buscando por entradas que representem tais pontos de fronteira (com valor *VERDADEIRO*). Para cada entrada *VERDADEIRA* **A_{ijk}** de **A**, nosso algoritmo computa o valor médio das cargas de cada

posição vizinha não nula de M_{ijk} . A entrada M_{ijk} recebe então o valor médio das cargas recém-calculado.

Supondo que a distribuição de cargas dentro de um átomo seja homogênea, a matriz resultante contém uma grade de cargas atômicas parciais nas posições correspondentes aos pontos de fronteira da matriz espacial, contida dentro dos limites da caixa. A complexidade da rotina de mapeamento é $O(m*n*o)$.

5.12 Rotinas Adicionais

Além das rotinas principais apresentadas no fluxograma e descritas anteriormente, nosso algoritmo realiza duas rotinas adicionais que permitem que um número maior de informações sobre os espaços vazios proteicos. Estas duas rotinas serão brevemente discutidas nesta seção.

5.12.1 A Rotina de Busca *a priori*

O objetivo da rotina de busca *a priori* é permitir que o programa encontre alvos de busca em uma proteína cujos sítios de ligação não são conhecidos. Ao executar esta busca, o usuário não precisa definir nenhum tipo de caixa. O KV-Finder cria, de maneira análoga às ferramentas existentes, uma caixa arbitrária e realiza uma busca global na proteína. Os resultados obtidos nesta busca servem como sugestão para o usuário estabelecer caixas e realizar outras buscas nessas regiões candidatas. Tendo em mente que o KV-Finder permite que o usuário ajuste os parâmetros da sua busca, os candidatos são ideais para iniciar o processo de ajuste da caixa até obter o detalhe de um espaço topológico interessante.

5.12.1.1 Detalhes da Implementação

Para realizar a busca *a priori*, nosso algoritmo define uma caixa arbitrária utilizando os extremos da proteína. Todos os átomos da proteína serão compreendidos dentro desta caixa pré-definida. A rotina de busca *a priori* ocorre após as etapas de preenchimento, no fluxo de execução do programa.

Quando o usuário opta por fazer uma busca *a priori*, nosso algoritmo define uma matriz tridimensional booleana auxiliar $\mathbf{A}'$. Essa matriz será preenchida utilizando a rotina padrão de preenchimento. Contudo, o preenchimento dessa grade auxiliar utilizará um raio de sonda maior do que a sonda padrão. Essa sonda auxiliar também pode ser definida como parâmetro de entrada e seu valor *default* é definido em 5.0 Å.

Após o preenchimento da matriz auxiliar, nossa rotina percorre a matriz auxiliar $\mathbf{A}'$ a procura de entradas de valor *VERDADEIRO*. Para cada entrada $\mathbf{A}_{ijk}'$ de valor *VERDADEIRO*, a entrada correspondente na matriz tridimensional padrão $\mathbf{A}$ recebe *FALSO*. Essa definição pode ser expressa na relação:

$$\mathbf{A}_{ijk} \text{ recebe } \textit{FALSO} \text{ para todo } \mathbf{A}_{ijk}' \textit{ VERDADEIRO},$$

para $0 \leq i < m$, $0 \leq j < n$ e $0 \leq k < o$.

Uma maneira intuitiva de compreender o que esta rotina realiza é pensar que excluímos do nosso modelo todos os pontos acessíveis a uma molécula esférica muito grande, obtendo assim apenas os pontos acessíveis a moléculas menores. O tempo de execução da rotina de busca *a priori* é limitado pela rotina de preenchimento e é da ordem de $O(m*n*o / h^3)$.

5.12.2 Identificação de resíduos expostos

Um recurso desejável em ferramentas como o KV-Finder para obter informações complementares sobre um espaço vazio proteico é identificar quais resíduos da proteína interagem com esse espaço. Portanto, implementamos uma rotina simples que exporta em um arquivo texto um código numérico identificando quais resíduos estão em contato com os espaços vazios da proteína.

5.12.2.1 Detalhes da Implementação

KV-Finder efetua a rotina de identificação de resíduos expostos após a rotina de preenchimento. Esta rotina consiste em percorrer o vetor de átomos e, para cada átomo do vetor, verificar se existe alguma entrada de valor *VERDADEIRO* na submatriz que contém este átomo. Caso essa entrada exista, o resíduo ao qual este átomo pertence é exportado para um arquivo de texto. A complexidade desta rotina é da mesma ordem da primeira etapa da rotina de preenchimento, $O(N/h^3)$, onde **N** é o número de átomos no vetor e **h** é o passo da matriz.

5.13 Criação de um modelo de saída dos dados obtidos

Após os espaços negativos serem propriamente mapeados, identificados e agrupados e o volume e a área de cada uma deles for calculado, é necessário exportar essas informações em um arquivo de saída. A maior preocupação ao gerar uma saída é

permitir que a informação gerada possa ser lida pela grande maioria dos softwares de modelagem existentes.

O formato de arquivo padrão empregado pelas ferramentas de modelagem para visualizar e manipular estruturas proteicas é o formato PDB. Como já mencionado, a informação estrutural em um arquivo PDB é representada pelas coordenadas do centro dos átomos que constituem uma dada proteína. Dada a sua alta acessibilidade e facilidade de uso, esse formato é adequado para conter a informação espacial gerado pelo nosso utilitário. Entretanto, o formato PDB não contém nenhum tipo de informação eletrostática.

Tendo isso em mente, nosso software produz dois arquivos de saída: um arquivo PDB e um arquivo PQR [25]. Nosso arquivo PDB de saída contém um átomo de hidrogênio centrado em cada entrada verdadeira da matriz booleana **A**. Os diferentes espaços vazios são representadas por cadeias diferentes e podem ser facilmente discriminadas em qualquer ferramenta de modelagem. O resultado é uma representação geométrica satisfatória do espaço mapeado.

Ao contrário das propriedades geométricas, não existe um formato largamente aceito para manipular e visualizar propriedades eletrostáticas de macromoléculas. Logo, decidimos empregar um formato não convencional chamado de PQR. Esse formato é uma extensão do formato PDB, que contém também o raio e a carga de todos os átomos da proteína. O formato de arquivo PQR permitiu que uma descrição detalhada do espaço compreendido dentro dos limites da caixa fosse computada, suprimindo assim as necessidades do nosso programa.

6. Resultados

A validação de nosso software foi realizada nos mais diversos tipos de espaços vazios encontrados em proteínas e enzimas com papéis catalíticos distintos. Enquanto algumas ferramentas se especializam em cavidades e outras em túneis, a maioria das fendas superficiais não são detectadas pelos utilitários existentes. Portanto, o uso da informação estrutural de uma maneira sistemática é severamente dificultado. Tendo isso em mente, decidimos comprovar as capacidades de caracterização de todos os tipos de vazio proteico do KV-Finder, mapeando e caracterizando túneis, bolsões, invaginações e fendas superficiais.

6.1 Descrição espacial e eletrostática de uma fenda superficial

A maioria das ferramentas de detecção de vazios existentes não apresenta bons resultados no tratamento de fendas, depressões e outros tipos de elementos topológicos superficiais. Contudo, muitos sítios de ligação localizam-se nas regiões mais superficiais das estruturas proteicas. Portanto a caracterização dessas fendas é essencial para agregar informação sobre proteínas que medeiam interações superficiais.

Para validar a capacidade do nosso software KV-Finder de mapear as propriedades espaciais e eletrostáticas de fendas superficiais, decidimos caracterizar uma depressão alongada encontrada na superfície da estrutura tridimensional de uma proteínas quinase. As proteínas quinase são uma família abundante de enzimas que realizam a fosforilação de diversas proteínas, geralmente resultando em mudanças funcionais importantes para

sinalização celular [26]. Essas enzimas apresentam um bolsão profundo no qual se liga um nucleosídeo trifosfato, geralmente o trifosfato de Adenosina (ATP), de cujo grupo funcional a ser transferido no processo de fosforilação será oriundo (Figura **20A**). Anexo a essa cavidade, existe uma formação chata que se estende por uma grande área da superfície proteica (Figura **20B**), apresentando múltiplos elementos topológicos e eletrostáticos que, combinados, regulam a especificidade dessa classe de enzimas para sua interação com as mais diversas proteínas alvo de fosforilação.

Figura 20: A cavidade do nucleotídeo (**A**) e a fenda superficial (**B**) presentes na estrutura tridimensional da Proteína Quinase A de camundongo (PDB ID: 1FMO), apresentadas com seus respectivos ligantes, uma molécula de ATP e o peptídeo inibidor de Welsh, também conhecido como PKI.

O diferencial de nosso programa que permite que a caracterização de fendas superficiais seja feita com maior precisão é a nossa abordagem direcionada, mediada por uma caixa tridimensional definida e customizada pelo usuário. Ao utilizar limites de busca

intuitivamente definidos ao invés de limites automáticos e inflexíveis, nós conseguimos solucionar o problema de discriminar entre os espaços vazios e o espaço externo a estrutura proteica. Além disso, esse método pode ser usado de maneira iterativa de modo a gerar mapas progressivamente melhores dos vazios proteicos.

Outro recurso introduzido pelo KV-Finder foi a capacidade de gerar um mapa parcial de cargas para descrever as propriedades eletrostáticas das superfícies encontradas, o que não é realizado por nenhum software existente com tamanho grau de detalhe e especificidade. A caracterização das propriedades espaciais e eletrostáticas da do sítio de ligação de uma proteína quinase cAMP-dependente (PKA) de *Mus musculus* obtida pelo KV-Finder está ilustrada na figura 21.

Figura 21: Mapeamento das propriedades eletrostáticas e espaciais da depressão superficial encontrada no sítio catalítico da proteína quinase dependente de cAMP de *Mus musculus* (**PDB ID:** 1FMO). Os resultados estão apresentados juntamente com a estrutura do peptídeo inibidor de Welsh, PKI(5-24). Duas argininas essenciais para a ancoragem do ligante estão destacadas em verde.

Após a obtenção dessa superfície, a estrutura de um dos ligantes dessa proteína

foi adicionada ao modelo para auxiliar na compreensão do caráter físico-químico da interação ligante-proteína. Conforme esperado, o ligante, o peptídeo inibidor de Walsh PKI (resíduos 5 a 24), apresenta grande complementariedade geométrica e eletrostática com a superfície do sítio de ligação obtida pelo KV-Finder.

Dois resíduos arginina (Figura **21**, em verde) são essenciais para a interação do ligante na Proteína Quinase A. Dado o alto pKa da arginina, esses resíduos apresentam cargas positivas em pH fisiológico, conferindo cargas básicas ao ligante. O mapeamento eletrostático da fenda superficial da PKA de revelou duas regiões vermelhas próximas a esses resíduos básicos, indicando cargas negativas complementares presentes na superfície da depressão.

Comparamos os resultados obtidos pelo KV-Finder na caracterização da fenda superficial da proteína quinase A murina com os resultados obtidos pelos utilitários Pocket Finder (Figura **22A**), Q-Site Finder (Figura **22B**) e CASTp (Figura **22C**). O Pocket Finder e o Q-Site Finder empregam a abordagem de matriz tridimensional e o CASTp emprega abordagem envolvendo forma-alfa.

A ferramenta Pocket Finder foi capaz de detectar e caracterizar apenas uma cavidade no interior da proteína, que corresponde ao bolsão do ATP e está representado em azul na figura **22A**. O software Q-Site Finder também detectou esta cavidade (ciano em **22B**) além de outros espaços menores sem grande relevância biológica. O software CASTp lida com problemas na caracterização da fenda superficial por conta da abordagem de forma-alfa, apresentando um resultado insatisfatório. As três ferramentas foram incapazes de detectar e caracterizar a fenda superficial de maneira precisa.

A diferença de qualidade das figuras geradas pelas outras ferramentas e a figura gerada pelo KV-Finder serve para ilustrar os benefícios obtidos pela integração do nosso programa em um software de modelagem poderoso. Isso permite ao usuário gerar figuras

em alta resolução, customizando a forma e as cores da figura, contraposto a interface gráfica dos softwares de visualização empregados por outros programas de detecção e caracterização de elementos topográficos na estrutura proteica.

A

B

C

D

Figura 22: Comparação entre a caracterização espacial de uma fenda superficial na estrutura de uma Proteína Quinase AMPc-dependente de camundongo, realizada por diferentes ferramentas: Pocket Finder (A), Q-Site Finder (B), CASTp (C) e a nossa ferramenta KV-Finder (D).

Além da fenda superficial, decidimos validar nossa ferramenta computando o volume da cavidade do nucleosídeo trifosfato que fornece o grupo fosfato durante o processo de fosforilação catalisado por essa família de enzimas. Na PKA descrita acima, o ligante desse bolsão é o trifosfato de adenosina (ATP). Devido a capacidade de particionamento do KV-Finder, nós pudemos isolar a cavidade do ATP e computar seu volume (Figura 23). O volume estimado para a cavidade do ATP foi de 343 Å³, empregando um passo de 0.3 Å. O volume da molécula de ATP, conforme descrito pelo ChemBank [27], é de 329 Å³. Além disso, o mapa parcial de cargas revelou um alto grau de complementariedade para com o ligante, sendo que a região complementar aos três fosfatos pode ser distinguida através de três regiões azuis (positivamente carregadas) na superfície encontrada.

Figura 23: Mapeamento das propriedades eletrostáticas e espaciais da cavidade do ATP encontrada no sítio catalítico da proteína quinase dependente de cAMP de *Mus musculus* (PDB ID: 1FMO). Três áreas azuis representam regiões positivamente carregadas da superfície proteica. Essas regiões são complementares ao trifosfato da molécula do ligante, ATP.

Comparamos então o resultado obtido com o resultado encontrado por outras ferramentas existentes. O melhor resultado encontrado foi obtido pela ferramenta Pocket Finder (Figura 24A), os outros resultados foram omitidos. Apesar de detalhar as propriedades espaciais da cavidade do ATP, o Pocket Finder apresenta um grau de detalhe inferior ao do KV-Finder, além de ser incapaz de realizar o mapeamento eletrostático.

Figura 24: Comparação entre a caracterização da cavidade do ATP da proteína quinase AMPC-dependente de camundongo, obtida pelo Pocket Finder (A) e pelo nosso utilitário KV-Finder (B).

6.2 Particionamento de um sítio ativo e integração em análise de larga escala

Uma limitação muito comum encontrada na maioria dos softwares que detectam e caracterizam espaços vazios em estruturas proteicas é a incapacidade de individualizar, e

portanto, descrever corretamente espaços intramoleculares adjacentes envolvidos em reações sequenciais de catálise. Como exemplo para validar esta funcionalidade em nossa ferramenta, validamos nosso algoritmo na enzima Aldeído desidrogenase (ALDH).

A super-família de enzimas Aldeído desidrogenase está envolvida na catálise de desidrogenização de um diverso número de aldeídos. Essas enzimas apresentam um sistema dual de cavidades, sendo que uma delas serve como sítio de ligação para o aldeído, sendo portanto variável dentre as diversas famílias, enquanto a outra serve como sítio de ligação para o cofator Dinucleotídeo Nicotinamida (NAD⁺).

Estas enzimas realizam três reações sequenciais: acilação, desidrogenização e tioacil hidrólise [28]. A catálise é complexa e se inicia na porção do espaço vazio em que se liga o aldeído. Nesta etapa inicial, ocorre um processo de acilação, em que um átomo de enxofre de uma cisteína catalítica presente na superfície da cavidade intramolecular realiza um ataque nucleofílico ao carbono do grupo carbonil do aldeído. Então, um grupo funcional hidreto é liberado do aldeído e atravessa um pequeno canal entre as duas porções da cavidade, sendo ultimamente transferido para o cofator, situado em um espaço intramolecular oposto ao do substrato. O intermediário tioacil sofre então uma oxirredução, gerando um ácido carboxílico que é o produto da reação.

Uma das motivações que levou a adição da funcionalidade de particionamento de sítios de ligação foi o interesse em calcular o volume de cada uma das porções da cavidade da aldeído desidrogenase separadamente, uma vez que apenas a porção da cavidade correspondente ao aldeído não apresenta alto grau de conservação. Nossa abordagem de caixa direcionada permite ao usuário dividir o sítio ativo calculando as propriedades espaciais e eletrostáticas apenas da porção de interesse. Dessa forma, o KV-Finder é capaz de discriminar entre as propriedades de diferentes espaços intramoleculares adjacentes.

As duas porções do sítio ativo da aldeído desidrogenase mitocondrial humana (ALDH2) foram separadas pelo nosso algoritmo, permitindo que o cálculo de volume de cada parte ocorresse de maneira independente (Figura 25). O volume computado da porção correspondente ao sítio de ligação do aldeído (Figura 25A), com um passo de 0.5 Å, foi de aproximadamente 402 Å. Já o volume da daidzina, um inibidor da ALDH2, é de 352 Å [27]. Analogamente, o volume calculado para o espaço correspondente ao sítio de ligação do cofator foi de 694 Å (figura 25B) e o volume do NAD⁺ é estimado em 453 Å [27].

Figura 25: Caracterização espacial do particionamento de espaços intramoleculares adjacentes da aldeído desidrogenase mitocondrial humana (**PDB ID: 2VLE**), conforme computado pelo KV-Finder. A porção que contém o sítio de ligação do aldeído está representada em **A**, sobreposta com o inibidor Daidzina. A porção que contém o sítio de ligação do cofator NAD⁺ sobreposta com seu ligante dinucleotídico está representada em **B**.

Para avaliar a variação de volume na porção do sítio catalítico que continha o

aldeído, foi realizada uma análise em larga escala empregando o KV-Finder para realizar o cálculo de volume. Este teste foi realizado com enzimas da super-família ALDH, selecionadas com base em dois critérios: apenas enzimas com disponibilidade da estrutura tridimensional no banco PDB e enzimas pertencentes as famílias ALDH1A e ALDH2 foram utilizadas.

Para cada proteína, foi realizado um alinhamento local com a sequência de aminoácidos das ALDHs humanas permitindo assim a classificação das proteínas alvo em 2 famílias: ALDH1A e ALDH2. Todas as estruturas foram sobrepostas espacialmente utilizando o software de modelagem YASARA [29,30,31]. Após a sobreposição, uma caixa foi definida delimitando a porção do sítio ativo de interesse, correspondente a região de ligação dos aldeídos, sendo então realizada a análise.

No total, foram avaliados os volumes dos espaços vazios de 57 enzimas. Nossos resultados demonstram que o volume da porção do aldeído dos sítios catalíticos das enzimas da família ALDH2 são significativamente menores que os mesmos volumes calculados para as enzimas da família ALDH1A (figura **26**). Um teste t de Welch rejeitou a hipótese nula (diferença nas médias igual a 0) a 1% de significância, com um p-valor de 2.067×10^{-7} .

Nossos resultados corroboram com as informações encontradas na literatura. As enzimas ALDH2 catalisam a transformação de acetaldeído em ácido acético, moléculas relativamente pequenas. Em contrapartida, a família ALDH1A catalisa a transformação do retinaldeído, uma molécula maior do que o ligante das ALDH2, em ácido retinóico.

Figura 26: Box plot dos volumes computados pelo KV-Finder da porção da cavidade que contém o sítio de ligação do aldeído de 57 proteínas da superfamília ALDH, classificadas em três famílias distintas. Os volumes das enzimas da família ALDH2 são significativamente menores que o da família ALDH1A, corroborando com os dados biológicos.

6.3 Alto grau de detalhe na descrição de espaços vazios

Alguns espaços vazios são formados durante mudanças conformacionais que ocorrem durante o mecanismo de ligação de um substrato na estrutura proteica. Estes espaços são complexos e sua detecção e caracterização é extremamente difícil, uma vez que esses vazios intramoleculares necessitam de um alto grau de detalhe para serem descritos de maneira correta.

Resolvemos validar nosso algoritmo utilizando um exemplo com grande importância econômica. A HIV aspartil-protease é uma proteína contida em um poligene viral, e seu papel é clivar a poliproteína resultante da expressão desse poligene produzindo assim as proteínas maduras essenciais para o ciclo de vida do HIV [32]. A HIV aspartil-protease apresenta duas subunidades diméricas que contêm uma estrutura em forma de aba (*flap*) na ponta do domínio. Conforme o substrato se aloca no sítio de ligação, essas abas se fecham, enclausurando o ligante e formando um túnel.

Como a HIV aspartil-protease é essencial para o ciclo de vida do vírus, esta enzima é alvo de terapêuticos e diversos compostos de nova geração podem e estão sendo desenvolvidos explorando seu sítio de ligação. Dessa forma, agrupar um conjunto maior de informações sobre este sítio é um processo importante para o desenho de fármacos mais eficientes e específicos.

Dada a complexidade e tamanho do túnel, para sua detecção eficiente é necessário utilizar um passo pequeno para caracterizar o canal formado. Contudo, uma limitação dos softwares existentes que usam abordagem matricial para a detecção de espaços vazios é o tamanho do passo, em que nessas ferramentas é restrito a valores maiores do que 0.9 Å.

KV-Finder é capaz de identificar e caracterizar o túnel formado pelos dois

monômeros da HIV aspartil-protease após o fechamento das abas sobre um inibidor chamado de MVT-101 [33]. Essa caracterização só é possível devido ao fato de que nossa ferramenta permite que o usuário defina valores arbitrariamente pequenos de passo. O mapeamento altamente detalhado do túnel foi obtido utilizando um passo de 0.1 Å (figura 27), aproximadamente uma ordem de grandeza menor do que aquele estabelecido pelas ferramentas existentes.

Figura 27: A superfície do canal da HIV Aspartil-protease (**PDB ID: 3FSM**) formado pela constrição das abas sobre o ligante gerada utilizando a nossa ferramenta. A representação da superfície do túnel, obtida com um passo de 0.1 Å, foi sobreposta ao inibidor MVT-101 (**B**).

Decidimos comparar este resultado, obtido pelo KV-Finder, com os obtidos por outras ferramentas existentes (figura 28). O melhor resultado foi encontrado utilizando o Pocket-Finder, portanto as imagens geradas pelas outras ferramentas foram omitidas. O grau de detalhe, e portanto, a precisão do KV-Finder são significativamente maiores do que o dos outros utilitários do gênero.

Figura 28: Comparação entre as caracterizações do túnel da HIV-aspartil protease efetuadas pelo software Pocket Finder (**A**) e pela nossa ferramenta, KV-Finder (**B**). Nossa ferramenta é capaz de alcançar um grau de detalhe muito maior que o das ferramentas existentes.

6.3.1 Ajuste dos espaços vazios a um conjunto de ligantes

Outra inovação apresentada pelo KV-Finder é a função de ajuste dos espaços vazios proteicos a um conjunto de ligantes fornecidos durante a entrada, destacando assim regiões que não estão dentro de uma distância de corte destes ligantes. Esta funcionalidade é particularmente útil quando se deseja visualizar quais regiões de um elemento topológico não estão sendo exploradas por um conjunto de drogas.

Para ilustrar este recurso, realizamos uma análise com um conjunto de nove

inibidores da HIV-aspartil protease, utilizando uma distância de corte de 2.5 Å (Figura 29). Esta distância foi escolhida por ser a menor distância necessária para acomodar um novo radical que poderia ser anexado aos ligantes existentes.

KV-Finder encontrou diversas regiões do túnel da HIV-aspartil protease que distam mais de 2.5 Å do conjunto de drogas fornecido como entrada. Estas regiões podem ser exploradas para aumentar a especificidade das drogas existentes, permitindo assim que os resultados gerados pelo KV-Finder sirvam como base para o desenho racional de fármacos.

Figura 29: Caracterização do túnel da HIV-aspartil protease após a rotina de ajuste ao ligante realizada pelo KV-Finder, apresentada com o peptídeo MVT-101 (**A**) e com um conjunto de nove inibidores da proteína (**B**). As regiões em vermelho são regiões que distam mais que 2.5 Å de todos os átomos do conjunto de ligantes fornecidos como entrada.

6.4 Particionamento de um Sítio de Ligação Complexo

Para finalizar a validação da nossa ferramenta, optamos por mapear os espaços vazios de uma proteína que apresentasse uma rede complexa de túneis alongados. Muitas vezes, ao tratar esse tipo de formação com uma ferramenta de detecção de cavidades, o usuário pode se deparar com uma quantidade exacerbada de informação que dificulta a obtenção de resultados.

Novamente, uma enzima com grande importância econômica foi escolhida para realizar a validação biológica do KV-Finder. A enzima conversora de angiotensina I (*human angiotensin I-converting enzyme* ou ACE) é um membro da família das carboxipeptidases e pode ser considerada como o principal alvo de drogas no tratamento da hipertensão e outras complicações cardiovasculares [34].

As enzimas da família das carboxipeptidases apresentam um sítio ativo relativamente grande, formado através da interação de um par de domínios homólogos ao redor de um zinco catalítico. Dado o caráter proteolítico da enzima, existe uma constrição nas proximidades do metal, acomodando as porções N- e C-terminal do peptídeo substrato em regiões opostas a ligação se clivagem S1-S1'. A rede intrincada de túneis se estende para cada uma dessas regiões como resultado da interação domínio-domínio.

As drogas disponíveis atualmente para o tratamento da hipertensão podem se ligar em qualquer uma das regiões ao redor do zinco catalítico. Contudo, esses fármacos são capazes de interagir apenas com uma parcela central e limitada do túnel proteico. Muitas das ramificações presentes na rede permanecem não exploradas pela grande maioria dos fármacos existentes.

KV-Finder foi utilizado para caracterizar a intrincada rede de túneis de uma

isoforma do gene *ACE*, a ACE testicular, que apresenta um domínio catalítico C-terminal idêntico ao da ACE somática. Através da nossa ferramenta, fomos capazes de identificar e caracterizar a rede complexa em sua totalidade (figura **30A**). Apresentamos também o resultado encontrado pelo Pocket Finder (figura **30B**) para ilustrar a comparação entre o resultado obtido pela nossa ferramenta e pelas outras ferramentas existentes.

Figura 30: Caracterização do túnel da Enzima Conversora de Angiotensina I humana, como computado pela nossa ferramenta KV-Finder (**A**) e pela ferramenta Pocket Finder (**B**). As ferramentas existentes são capazes de gerar uma caracterização do túnel completo da enzima.

A estrutura de uma droga conhecida por ocupar o maior espaço do sítio ativo, o inibidor fosfínico altamente C-domínio específico RXPA380, foi sobreposta na estrutura encontrada ocupando apenas uma parcela pequena do túnel encontrado. Portanto, apesar de termos caracterizado o túnel da ACE, informações sobre o volume e a área do túnel completo não são tão significativas, uma vez que apenas uma pequena região do túnel é ocupada pelos inibidores existentes. Seria desejável poder segmentar o túnel,

computando apenas as propriedades físico-químicas de uma região de interesse.

Mais uma funcionalidade exclusiva do KV-Finder é a sua habilidade exclusiva de seccionar os espaços encontrados, calculando o volume apenas de uma região específica definida pela caixa de busca. Para validar este recurso, particionamos o túnel alongado da enzima ACE. Dessa forma, fomos capazes de calcular o volume e a área apenas da região do túnel na qual interagem os terapêuticos existentes (Figura 31).

Figura 31: O túnel da ACE computado pelo KV-Finder, dessa vez com os limites da caixa definidos ao redor do inibidor fosfínico RXPA380. Nossa abordagem de caixa direcionada pelo usuário permite que um túnel muito extenso seja seccionado, facilitando o cálculo das propriedades volumétricas e eletrostáticas de uma região de interesse.

Ao observarmos a caracterização gerada pelo KV-Finder dessa região específica, notamos a presença de pequenas protuberâncias nas bordas do túnel. Tendo em mente que a ACE é uma peptidase, estas protuberâncias poderiam estar associadas com as cadeias laterais dos aminoácidos de um peptídeo a ser clivado pela enzima. Portanto,

decidimos subseccionar o túnel da ACE, próximo a região S1', estimando o volume destas protuberâncias (Figura 32).

Figura 32: Pequenos bolsões encontrados no túnel da enzima ACE foram caracterizados graças à capacidade exclusiva do KV-Finder de segmentar e subsegmentar regiões de busca na estrutura proteica.

O volume de cada um desses pequenos bolsões pôde ser estimado, fornecendo informações importantes para guiar o desenho de um novo inibidor possivelmente mais específico para a enzima. Isso é possível uma vez que estes volumes podem ser utilizados para escolher grupos funcionais candidatos de volume similar aos computados para cada bolsão. Estes grupos funcionais candidatos poderiam ser anexados a um inibidor aumentando a sua especificidade.

Além disso, os resultados revelaram que essas subsecções apresentam, de fato, volumes relativamente próximos ao volume de aminoácidos, mais uma vez demonstrando a alta complementariedade espacial entre sítio de ligação e ligante. Realizamos então uma comparação entre o volume de cada uma destas secções com o volume dos 20

aminoácidos canônicos. Isso permitiu a inferência de aminoácidos putativos que poderiam compôr o peptídeo substrato original e que seriam interessantes candidatos para serem anexados a um fármaco (Tabela 1). Este método poderia ser usado para prever possíveis sequências de clivagem, baseadas nas informações volumétricas e eletrostáticas das protuberâncias anexas ao túnel de carboxipeptidases.

	Volume (Angstrom ³)	Aminoácido mais próximo	Volume do Aminoácido (Angstrom ³)
Área Vermelha	139.848	Histidina	137.14
Área Verde	175.104	Tirosina	163.18
Área Azul	58.640	Glicina	57.05
Área Amarela	134.656	Metionina	131.19
Área Magenta	71.232	Alanine	71.09

Tabela 1: Volume das protuberâncias encontradas nas proximidades da região S1' do túnel da enzima ACE. Os bolsões apresentam volumes muito similares a aminoácidos canônicos, o que indica que essas protuberâncias podem acomodar cadeias laterais durante o processo de clivagem de peptídeos.

6.5 Análise de Precisão e Tempo de Execução

Realizamos uma análise da precisão e tempo de execução de nossa ferramenta para diferentes tamanhos de passo para um conjunto de nove proteínas. Esta análise foi realizada em um computador Intel Core i5 M450, com 4GB de memória e sistema operacional Ubuntu 10.04, Kernel linux 2.6.32.29-generic. Para realizar uma análise comparativa, empregamos três das ferramentas mais eficientes para detecção e caracterização de espaços vazios na superfície proteica: o Pocket Finder, o CASTp e o fPocket.

Para ilustrar a capacidade do nosso método de analisar virtualmente qualquer tipo de espaço vazio na estrutura proteica, analisamos nove espaços vazios com ligantes conhecidos, comparando o volume encontrado pelo nosso software com o volume dos ligantes e com as outras ferramentas existentes (figura **33**). Através dessa análise, observamos que os resultados encontrados pelo KV-Finder são relativamente próximos aos volumes dos ligantes. Já as outras ferramentas existentes não foram capazes de identificar e caracterizar alguns espaços negativos, e apresentaram resultados ruins comparados ao volume de alguns ligantes. Dessa forma, KV-Finder foi a única ferramenta capaz de obter bons resultados em todas as proteínas analisadas.

Realizamos então uma análise de convergência do cálculo de volume, variando o tamanho do passo e avaliando a variação do volume para o nosso conjunto de nove proteínas. Observamos que para todas as proteínas encontradas o volume converge a partir de um passo de 0.2 Å (exemplificado através da figura **34**).

Figura 33: Volumes encontrados por diversas ferramentas no cálculo de nove espaços vazios diferentes em nove proteínas. Observe que o volume encontrado pelo KV-Finder (laranja) está sempre próximo, porém um pouco maior que o volume do ligante (azul). As outras ferramentas existentes não foram capazes de calcular as propriedades de alguns espaços vazios e apresentaram resultados ruins em outros casos.

Figura 34: Análise de convergência do volume da cavidade de uma proteína que se liga a ácidos graxos (*Fat Acid Binding Protein* ou *FABp*). O volume converge conforme o passo aproxima-se de 0.2 Å. Esta convergência foi encontrada em todas as proteínas do nosso conjunto de teste, sendo que a FABp foi selecionada para representar este resultado.

Realizamos também uma análise comparativa do tempo de execução de nosso software com diferentes tamanhos de passo comparados com as outras ferramentas do gênero (tabela 2). Estes dados demonstram que, com relação as ferramentas existentes, o KV-Finder apresenta um ganho de performance, obtido pela restrição do espaço de busca.

Ferramenta	Tempo de Execução
KV-Finder Passo 1.0 Å	<0,1 segundos
KV-Finder Passo 0.5 Å	Aproximadamente 0,1 segundos
KV-Finder Passo 0.4 Å	Aproximadamente 0,2 segundos
KV-Finder Passo 0.3 Å	Aproximadamente 0,5 segundo
KV-Finder Passo 0.2 Å	1-3 segundos
KV-Finder Função FULL PROTEIN	2-10 segundos
Q-Site Finder	5-60 segundos
Pocket Finder	5-60 segundos
fPocket	2-5 segundos

Tabela 2: Tempo de execução do KV-Finder com diferentes tamanhos de passo, comparados com outras ferramentas existentes para detecção de espaços vazios na estrutura proteica. Nossa ferramenta apresenta um ganho de eficiência ao restringir o espaço de busca.

7. Discussão

A estrutura tridimensional de proteínas abriga uma série de elementos topológicos e espaços vazios que contêm informações funcionais essenciais para compreender interações proteicas. A interação proteína-substrato é altamente específica e, portanto, exige um alto grau de complementariedade entre as propriedades espaciais e eletrostáticas da macromolécula e seu ligante. Dessa forma, é essencial agrupar o máximo de informações sobre o sítio de ligação para entender os mecanismos que regulam essa interação.

A caracterização espacial e eletrostática de espaços vazios na estrutura tridimensional proteica é capaz de auxiliar a caracterização funcional de proteínas, bem como guiar o desenho racional de fármacos, possivelmente aumentando a especificidade dos terapêuticos existentes. Dessa forma, diversas abordagens computacionais foram desenvolvidas durante as últimas duas décadas de modo a tratar este problema . Contudo, as ferramentas existentes para busca e caracterização de cavidades na estrutura proteica lidam com uma série de limitações como a falta de um tratamento eletrostático apropriado, a falta de capacidade de particionamento dos espaços encontrados e a falta de precisão, especialmente no tratamento de elementos superficiais.

Dessa forma, desenvolvemos uma nova ferramenta, com o intuito de estender as funcionalidades das ferramentas existentes, fornecendo caracterizações mais detalhadas e informativas sobre os vazios proteicos. Através de uma abordagem inovadora que permite que o usuário defina uma caixa flexível em uma região de interesse, nossa ferramenta é capaz de determinar a forma e a extensão de espaços vazios e calcular o volume e a área desses vazios.

7.1 Estudo do Método

Conforme discutido na seção 2, existem dois métodos principais para o tratamento do problema de identificação e caracterização das propriedades espaciais de espaços vazios, as abordagens de matriz tridimensional e as abordagens envolvendo formas-alfa.

Tendo em mente que as abordagens de forma-alfa não consideram raio atômico, esses métodos não conseguem retratar os modelos de superfícies descritos na seção 5.6.1. Dessa forma, a superfície encontrada por abordagens forma-alfa é uma aproximação para a superfície real proteica. Portanto, em comparação com as abordagens envolvendo formas-alfa, as abordagens de matriz tridimensional apresentam uma maior precisão e um maior grau de detalhe ao caracterizar elementos topográficos.

Os métodos de forma-alfa dificultam o particionamento de elementos topográficos, uma vez que o conceito de forma-alfa define apenas a forma de um grupo de pontos, sem se preocupar em definir espaços vazios. Como nesta abordagem os espaços vazios são representados por tetraedros, um outro modelo precisaria ser desenvolvido para permitir a divisão destes tetraedros, bem como a decisão de quais tetraedros deveriam ser divididos de modo a permitir um particionamento adequado.

A busca em regiões específicas também é impossibilitada pela própria definição de forma-alfa. Caso um dos pontos seja retirado do conjunto, a forma encontrada não mais corresponde ao espaço proteico, sendo portanto obrigatório o uso de todos os átomos de uma proteína para a execução do algoritmo.

Por estas razões, optamos por implementar nossa ferramenta utilizando um método de matriz tridimensional ajustado, que permite um direcionamento realizado pelo usuário. Apesar desta decisão acarretar em uma pequena perda de eficiência, as

funcionalidades que julgávamos ausentes nos utilitários existentes poderiam ser melhor implementadas empregando este tipo de abordagem.

7.2 Tratamento de Elementos topográficos Superficiais

Devido ao problema de distinção entre o que é espaço externo a proteína e o que é elemento topográfico, as ferramentas de detecção de vazios existentes não apresentam bons resultados no tratamento de fendas, depressões e outros tipos de elementos topológicos superficiais. Por exemplo, a fenda superficial de proteínas quinase não é identificada por nenhuma ferramenta existente. Esta perda de informação é extremamente indesejável, uma vez que muitos sítios de ligação estão contidos em elementos superficiais.

Ao outorgar a responsabilidade da decisão do que é espaço externo e o que é elemento topológico ao pesquisador, nosso método direcionado permite que fendas superficiais sejam descritas de maneira eficiente e precisa, conforme apresentado na seção 6.1.

Além disso, nossa ferramenta apresenta um função para encontrar elementos topográficos sem nenhum conhecimento prévio da função proteica. Essa função fornece sugestões para o usuário, que pode então definir caixas em volta destas regiões, obtendo mapeamentos mais detalhados. Dado o alto nível de customização e o baixo tempo de execução, nossa ferramenta permite que o usuário gere sucessivas caracterizações, ajustando os limites da caixa, passo da matriz, tipo de superfície a ser calculada e parâmetros de filtragem a cada caracterização. Dessa forma, o usuário consegue ajustar

a caracterização de modo a definir o elemento topográfico de interesse de maneira precisa.

Os outros métodos de abordagem matricial (VOIDOO¹, POCKET⁸, LIGSITE⁹, Pocket Finder¹⁰, Q-Site Finder¹¹) apresentam dificuldades em definir a forma dos espaços topológicos superficiais, uma vez que empregam uma caixa arbitrária, fixa e definida pelo computador. A distinção entre espaço externo e um elemento topográfico é uma tarefa relativamente complexa para o computador, mas simples para um pesquisador com um pouco de prática e entendimento do seu caso de estudo.

As abordagens de forma-alfa^{12,13,14,35} acabam por excluir as fendas superficiais, uma vez que a aproximação dos átomos por pontos faz com que as superfícies encontradas por esse método não considerem raio atômico ou a sonda necessária para construção das superfícies de acesso ao solvente e molecular.

7.3 Caracterização Eletrostática

Obter uma caracterização eletrostática dos espaços mapeados é extremamente interessante para prever ou desenhar ligantes de uma proteína. Dada a alta complementaridade entre sítio ativo e ligante, observar regiões com concentrações de carga em um elemento topológico fornece informações cruciais para guiar o desenho racional de fármacos e a caracterização funcional.

Infelizmente, nenhuma das ferramentas existentes apresenta um recurso que permite que o usuário agregue informações eletrostáticas aos elementos topográficos identificados e caracterizados. Tendo isso em mente, implementamos em nosso software

uma função de descrição eletrostática da superfície, feita através de um mapa de cargas parciais. Este recurso exclusivo permite que o usuário agregue informações físico-químicas às superfícies computadas (conforme ilustrado na Seção 6.1), tornando as caracterizações realizadas pelo KV-Finder muito mais ricas e informativas.

O tratamento eletrostático realizado pelo KV-Finder é uma aproximação para o comportamento das cargas na superfície proteica. Certamente, uma abordagem mais robusta, envolvendo cálculos de campo de força, poderia ser empregada para descrever estas superfícies. Entretanto, esta abordagem acarretaria em uma perda significativa de eficiência. Portanto, optamos por utilizar a aproximação das cargas parciais para permitir que o cálculo das propriedades eletrostáticas possa ser realizado em um tempo de execução razoável.

7.3 Particionamento de Elementos Topológicos

Elementos topológicos podem apresentar regiões específicas que alocam um determinado ligante, seções contidas dentro do elemento que geralmente são identificadas e caracterizadas como um todo. Seria interessante permitir que o usuário pudesse calcular as propriedades geométricas de uma região específica, como um bolsão que surge em um túnel por exemplo, em vez de calcular apenas o volume total de um dado elemento topográfico.

O método direcionado pelo usuário empregado pelo KV-Finder introduz a função de particionar e subparticionar os espaços intramoleculares, permitindo que o usuário calcule o volume e a área de uma partição do espaço alvo, conforme demonstrado nas seções

6.2 e 6.4. Esta função aumenta o grau de informação gerada sobre um espaço vazio na estrutura tridimensional proteica e permite identificar grupos funcionais e cadeias laterais baseados em cálculos volumétricos de porções de um sítio ativo.

Os outros métodos matriciais (VOIDOO¹, POCKET⁸, LIGSITE⁹, Pocket Finder¹⁰, Q-Site Finder¹¹) não permitem que o usuário calcule o volume de regiões específicas de um túnel alongado, por exemplo, diminuindo o nível de informação gerada sobre o elemento topográfico. Quanto aos métodos de forma-alfa, a função de particionamento é inviável uma vez que este tipo de abordagem requer todos os átomos da proteína para definir uma forma aceitável.

Dessa forma, nossa ferramenta apresenta mais um recurso que estende as funcionalidades dos utilitários existentes, contribuindo para gerar um maior nível de informação geométrica sobre os espaços vazios caracterizados na estrutura proteica.

7.4 Maior Precisão no Cálculo de Volume

Conforme substantiado pela nossa análise de precisão (seção 6.5), os métodos existentes lidam com problemas na determinação do volume dos elementos topográficos encontrados. Apesar de encontrar bons resultados em algumas proteínas, outras proteínas apresentam resultados ruins. Em outras palavras, as ferramentas existentes apresentam uma precisão inconstante.

No caso de métodos matriciais utilizados para o cálculo de volume, a precisão é inversamente proporcional ao tamanho do passo da matriz. Como observado em nossa análise de precisão, o volume calculado através da integração numérica converge para

valores de passo próximos a 0.2 Å. A maioria das ferramentas que emprega matriz tridimensional booleana restringe o passo para valores maiores que 1.0 Å, apresentando então uma precisão limitada.

Nossa abordagem permite que os usuários definam passos arbitrariamente pequenos, limitados apenas pelas capacidades de hardware do usuário. Dessa forma, o KV-Finder é capaz de realizar o cálculo de volume com um passo significativamente menor do que as outras abordagens matriciais.

Através das comparações apresentadas na seção 6.5, fomos capazes de comprovar que o nosso software apresenta um ganho de precisão no cálculo de volume, comparado com outros utilitários do gênero (fPocket, Pocket Finder, CASTp), uma vez que o passo empregado é menor do que as outras ferramentas existentes. Mesmo comparado com as abordagens de forma-alfa, o KV-Finder foi capaz de calcular o volume dos espaços vazios com maior precisão, apresentando resultantes constantes ao longo dos casos de estudo.

Como o passo da matriz é inversamente proporcional à precisão no cálculo de volume e de área (ou seja quanto menor o passo, maior a precisão), a precisão da nossa ferramenta é dependente das capacidades do hardware de seus usuários, o que pode ser visto como uma limitação de nosso método.

7.5 Integração com Softwares de Visualização

KV-Finder foi integrado com uma ferramenta padrão de visualização molecular, o VMD, sendo fácil de usar e permitindo que pessoas sem experiência com programação gerem mapeamentos customizados de maneira simples e rápida. Uma das limitações do KV-Finder é que esta integração poderia ser otimizada. Infelizmente, o usuário ainda precisa importar o script dentro do console do VMD, o que para alguns usuários menos familiarizados com programação representa uma dificuldade. Seria ideal que o KV-Finder já viesse incorporado como uma rotina do próprio VMD e que houvesse um botão dentro da interface gráfica do VMD para utilização do KV-Finder. Isso facilitaria ainda mais o uso da nossa ferramenta. Um dos nossos planos futuros é contatar os desenvolvedores do VMD para incorporar nosso utilitário ao software de visualização molecular.

Outra limitação de nossa ferramenta é que, diferente de outras ferramentas existentes, o KV-Finder ainda não apresenta uma abordagem Web. Seria extremamente útil permitir que os usuários realizassem sua caracterização totalmente online, sem depender da instalação ou download do software. A implementação de uma abordagem Web para o KV-Finder também faz parte dos nossos planos futuros.

7.6 Considerações Finais

Em comparação com outros softwares existentes, nossa ferramenta resolve uma série de limitações permitindo o aumento de precisão, maior grau de customização,

capacidade de integração em larga escala, capacidade de particionamento de sítios ativos, capacidade de detectar fendas superficiais e de realizar um mapeamento eletrostático. Além disso, KV-Finder é capaz de analisar qualquer tipo de vazio encontrado em macromoléculas biológicas, inclusive os elementos superficiais..

Nossa ferramenta foi validada em diversas proteínas e enzimas, descrevendo os mais diversos tipos de espaços vazios em sítios de ligação e sítios catalíticos com as mais diferenciadas funções. A análise em larga escala revelou resultados que corroboram com a literatura, mais uma vez demonstrando o potencial de nossa ferramenta.

8. Conclusões

Através da validação apresentada nos resultados, nosso software foi capaz de:

- Identificar e caracterizar espaços vazios em regiões específicas na estrutura 3D de proteínas.
- Calcular o volume, a área da superfície e um mapa de cargas parciais desses elementos topológicos.
- Realizar análises em larga escala.
- Particionar sítios ativos e calcular as propriedades de cada seção.
- Ajustar cavidades a um ligante fornecido.
- Identificar virtualmente qualquer tipo de cavidade.

Dessa forma, KV-Finder fornece informações importantes para auxiliar a caracterização funcional de proteínas de papel biológico desconhecido, bem como direcionar o desenho racional de fármacos.

9. Bibliografia

- [1] Kleywegt, G.J. and Jones, T.A.(1997). Template convolution to enhance or detect structural features in macromolecular electron-density maps. *Acta Crystallogr D Biol Crystallogr.* 1997 Mar 1;53(Pt 2):179-85.
- [2] Bohacek, R.S. and McMartin, C. (1997). Modern computational chemistry and drug discovery: structure generating programs. *Curr Opin Chem Biol.* 1997 Aug;1(2):157-61.
- [3] Mattos C, Ringe D (1996). Locating and characterizing binding sites on proteins. *Nat Biotechnol.* 1996 May;14(5):595-9.
- [4] Nelson, D. L. and Cox, M. M. (2005) *Lehninger's Principles of Biochemistry*, 4th Edition, W. H. Freeman and Company, New York.
- [5] Helliwell JR (2005). "Protein crystal perfection and its application". *Acta Crystallogr. D Biol. Crystallogr.* 61 (Pt 6): 793.
- [6] Martin, G.E; Zekter, A.S., *Two-Dimensional NMR Methods for Establishing Molecular Connectivity*; VCH Publishers, Inc: New York, 1988 (p.59)
- [7] F.C.Bernstein, T.F.Koetzle, G.J.Williams, E.E.Meyer Jr., M.D.Brice, J.R.Rodgers, O.Kennard, T.Shimanouchi, M.Tasumi, "The Protein Data Bank: A Computer-based Archival File For Macromolecular Structures," *J. of. Mol. Biol.*, 112 (1977): 535.

- [8] Levitt DG, Banaszak LJ (1992). POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J Mol Graph*. 1992 Dec; 10 (4): 229 – 234.
- [9] Hendlich M, Rippmann F, Barnickel G. (1997). LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graph Model*. 1997 Dec;15(6):359-63, 389.
- [10] An J., Totrov M., Abagyan R.. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol Cell Proteomics*. 2005 Jun;4(6):752-61. Epub 2005 Mar 9.
- [11] Laurie ATR, Jackson RM (2005). Q-SiteFinder: an energy-based method for the prediction of protein–ligand binding sites. *Bioinformatics* 2005 21(9):1908-1916
- [12] Edelsbrunner H, Mücke E P (1994). Three-dimensional Alpha Shapes. *ACM Transactions on Graphics*. 13(1): 43 – 72.
- [13] Liang J, Edelsbrunner H, Woodward C (1998). Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci*. 1998 Sep;7(9):1884-97.
- [14] Joe Dundas, Zheng Ouyang, Jeffery Tseng, Andrew Binkowski, Yaron Turpaz, and Jie Liang. 2006. CASTp: computed atlas of surface topography of proteins with structural and

topographical mapping of functionally annotated residues. Nucl. Acids Res., 34:W116-W118.

[15] Laskowski RA (1995). SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph. 1995 Oct;13(5):323-30, 307-8.

[16] Smart OS, Neduvellil JG, Wang X, Wallace BA, Sansom MS. (1996). HOLE: a program for the analysis of the pore dimensions of ion channel structural models. J Mol Graph. 1996 Dec;14(6):354-60, 376.

[17] Website: <http://www.tcl.tk/>

[18] Humphrey, W., Dalke, A. and Schulten, K., (1996) "VMD - Visual Molecular Dynamics", J. Molec. Graphics, 1996, vol. 14, pp. 33-38

[19] DeLano, W.L. The PyMOL Molecular Graphics System (2002) on World Wide Web. <http://www.pymol.org>

[20] Case, D.A. et al. (2006). Amber 9. University of California. U. o. California. California

[21] Lee, B. and Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol., 1971. 55, 379-400

[22] Connolly, ML (1983). Solvent-accessible surfaces of proteins and nucleic acids. Science. 1983 Aug 19;221(4612):709-13.

- [23] Knuth, D. E. (1997), The Art Of Computer Programming Vol 1. 3rd ed., Boston: Addison-Wesley, ISBN 0-201-89683-4, OCLC 155842391
- [24] Cormen, T.H., Leiserson, C.E., Rivest, R. L. and Stein, C. (2001). Introduction to Algorithms, Second Edition. MIT Press and McGraw-Hill, 2001. ISBN 0-262-03293-7. Section 22.3: Depth-first search, pp. 540–549.
- [25] Website: <http://www.poissonboltzmann.org/file-formats/biomolecular-structure/pqr>
- [26] Lavogina, D., Enkvist, E., Uri, A., Engh, R.A. and Bossemeyer, D. (2010). Bisubstrate inhibitors of protein kinases: from principle to practical applications. ChemMedChem. 2010 Jan;5(1):23-34.
- [27] Website: <http://chembank.broadinstitute.org/welcome.htm>
- [28] Wymore, T., Deerfield II, D. W., Hempel, J. (2007). Mechanistic Implications of the Cysteine-Nicotinamide Adduct in Aldehyde Dehydrogenase Based on Quantum Mechanical/Molecular Mechanical Simulations. Biochemistry 2007, 46, 9495-9506
- [29] Krieger E., Koraimann G., Vriend G. (2002). Increasing the precision of comparative models with YASARA NOVA—a self-parameterizing force field. Proteins 47 (3): 393–402. doi:10.1002/prot.10104. PMID 11948792.
- [30] Krieger E., Vriend G. (2002). Models@Home: distributed computing in bioinformatics

using a screensaver based approach. *Bioinformatics* 18 (2): 315–8.

doi:10.1093/bioinformatics/18.2.315. PMID 11847079.

[31] Krieger E, Darden T, Nabuurs SB, Finkelstein A, Vriend G (2004). Making optimal use of empirical energy functions: force-field parameterization in crystal space. *Proteins* 57 (4): 678–83. doi:10.1002/prot.20251. PMID 15390263

[32] Lefebvre, E., Schiffer, C.A. (2008). Resilience to resistance of HIV-1 protease inhibitors: profile of darunavir. *AIDS Rev.* 2008 Jul-Sep;10(3):131-42.

[33] Johnson, E.C. et al (2007) Insights from atomic-resolution X-ray structures of chemically synthesized HIV-1 protease in complex with inhibitors. *Mol Biol.* 2007 Oct 26;37(3):573-86. Epub 2007 Aug 2.

[34] Corradi, H.R. et al (2007). The structure of testis angiotensin-converting enzyme in complex with the C domain-specific inhibitor RXPA380. *Biochemistry.* 2007 May 8;46(18):5473-8. Epub 2007 Apr 18.

[35] Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J. (2006) CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res.* 2006 Jul 1;34(Web Server issue):W116-8.

[36] Sonavane S, Chakrabarti P. (2010) Prediction of active site cleft using support vector machines. *J Chem Inf Model.* 2010 Dec 27;50(12):2266-73. Epub 2010 Nov 16.