PEDRO MAGALHÃES MARTINS

CAPRI: UMA BASE DE DADOS PARA ANÁLISE COMPARATIVA DE PARADIGMAS PARA PROSPECÇÃO DE CONTATOS EM INTERFACES PROTEÍNA-PROTEÍNA

Belo Horizonte

Agosto de 2015

PEDRO MAGALHÃES MARTINS

CAPRI: UMA BASE DE DADOS PARA ANÁLISE COMPARATIVA DE PARADIGMAS PARA PROSPECÇÃO DE CONTATOS EM INTERFACES PROTEÍNA-PROTEÍNA

Dissertação apresentada ao Programa de Pós-Graduação em Bioinformática do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Bioinformática.

ORIENTADOR: PROFA. DRA. RAQUEL CARDOSO DE MELO MINARDI
COORIENTADOR: PROF. DR. VINÍCIUS DINIZ MAYRINK

Belo Horizonte

Agosto de 2015

Agradecimentos

Agradeço primeiramente à Deus, pela vida, pela saúde e pela família.

Agradeço aos meus pais, por sempre estarem comigo e me apoiaram, me incentivando e me ajudando nos momentos difíceis. Agradeço também ao meu irmão por compartilhar a infância comigo, e todas as brincadeiras e aventuras vivadas.

Agradeço às minhas primas, Sheila e Fernanda, que sempre foram como mães para mim e sempre estiveram ao meu lado.

Agradeço aos meus amigos de laboratório que proporcionaram muitas felicidades e sorrisos nestes últimos anos. Agradeço ao Alexandre, pela ajuda e conhecimento passado sobre visualizações de dados. Agradeço ao Laerte, pelas excelentes linhas de código compartilhadas por e-mail. Agradeço à Larissa, por me ajudar a utilizar a ferramenta LaTeX ao qual pude confeccionar este trabalho. Agradeço à Valdete e à Sabrina pelos conselhos e pela disposição para tirar muitas dúvidas. Agradeço também aos companheiros de congressos: João, Wellisson, Sandro, Elisa e Kato.

Agradeço aos professores Carlos Henrique da Silveira (UNIFEI), Vinicius Diniz Mayrink (UFMG) e ao Leonardo Henrique Franca de Lima (UFSJ) pela grande contribuição para a conclusão deste trabalho.

Faço aqui também um agradecimento muito especial ao Prof. Marcelo Santoro, que infelizmente não está mais entre nós, mas que deixou muitas coisas boas que se perpetuam através da Prof.ª Raquel Melo Minardi, da Prof.ª Valdete e do Prof. Carlos Henrique da Silveira.

Termino aqui fazendo o meu segundo agradecimento especial à minha orientadora, Prof.ª Raquel Melo Minardi por todo apoio e dedicação dado para conclusão deste trabalho. Obrigado por sempre me incentivar e motivar, com todas as ideias e sugestões concedidas, sempre com muito entusiasmo e alegria. Obrigado pro ser também essa pessoal iluminada e espero que possam ter mais pessoas como você no meio acadêmico.

Resumo

Estudos que envolvem estrutura de proteínas lidam na maioria das vezes com uma grande quantidade de informação. Para compreender melhor o processo de interações proteína-proteína é necessário estudar os fenômenos que ocorrem em suas interfaces moleculares. Interações podem ser observadas do ponto de vista de resíduos, porém sabe-se que elas ocorrem na realidade em nível atômico. Vários paradigmas propõem formas distintas para definir interações atômicas e sabe-se que é necessário comparar estes paradigmas para melhorar nossa compreensão sobre interações moleculares, permitindo abranger nosso conhecimento sobre os vários mecanismos e funções celulares. Com isso, propomos aqui o banco de dados CAPRI, para análise comparativa entre três paradigmas distintos que são usados para definir contatos em interface proteína-proteína. O banco de dados CAPRI possui informações de cerca de 45 mil complexos proteicos, contendo dados quanto as interações realizadas entre pares de átomos, resíduos e cadeias. Ao todo, quatro tipos de interações são investigadas, sendo elas: pontes de hidrogênio, interações hidrofóbicas, pontes salinas e empilhamento aromático. Os resultados obtidos, juntamente com a obtenção da base de dados criada podem ser acessados através do endereço: http://homepages.dcc.ufmg.br/~pmartins/capri1/.

Palavras-chave: Contatos atômicos, Interação proteína-proteína, Interface, Estrutura de proteína, Base de dados, Delaunay.

Abstract

Studies involving protein structures most often deal with a large amount of information. To understand the process of protein-protein interactions is necessary to study the processes that occur in their molecular interfaces. Interactions can be observed at the residue level, but it is known that they occur in reality at the atomic level. Several paradigms propose different ways to define atomic interactions, and it is known that it is necessary to compare these paradigms to improve our understanding of the molecular interactions allowing to expand our knowledge of the many mechanisms and cellular functions. Thus, we propose here the CAPRI database for comparative analysis of three different paradigms that are used to define contacts in protein-protein interface. CAPRI has information of about 45,000 protein complexes containing data about interactions between pairs of atoms, residue and chains. Four types of interactions are investigated: hydrogen bonds, hydrophobic interactions, salt bridges and aromatic stacking. The results and the database can be accessed through the link: http://homepages.dcc.ufmg.br/~pmartins/capri1/.

Keywords: Atomic contact, Protein-protein interaction, Interface, Protein structure, Database, Delaunay.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

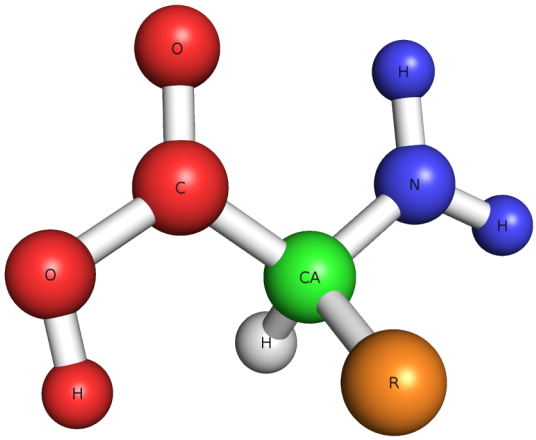
1.1 Proteínas

As proteínas são macromoléculas de grande importância para a existência dos seres vivos. Compostas por unidades menores chamadas de aminoácidos, as proteínas desempenham papéis fundamentais nos organismos vivos, atuando como transportadoras de oxigênio, no caso das hemoglobinas; reguladoras de funções corporais, onde se incluem os hormônios; proteção inumológica; catalisadoras em reações bioquímicas no processo de metabolismo, entre outras infinidades de funções fundamentais [Stryer et al., 2004]. Em uma única célula encontramos uma grande variedade de proteínas; milhares de diferentes tipos e com funções distintas [Nelson & Cox, 2014].

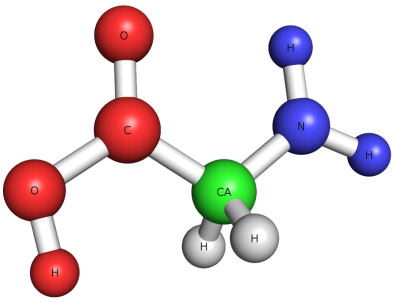
1.1.1 Aminoácidos

Os aminoácidos são pequenas moléculas fundamentais que compõem as proteínas. São unidades estruturais básicas comumente chamados de resíduos de aminoácidos, ou simplesmente resíduos, quando se unem em ligações peptídicas, formando a cadeia polipeptídica. O uso do termo "resíduo"se deve a perda de átomos para composição de um molécula de água que é liberada quando um aminoácido se une a outro [Nelson & Cox, 2014].

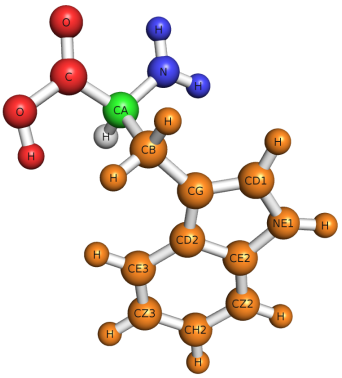
Ao todo existem 20 aminoácidos comumente encontrados nos seres vivos. São formados por um átomo central de carbono (denominado carbono-α) ligado a um grupamento amina ($-NH_2$), uma carboxila ($-COOH$), um átomo de hidrogênio e um grupo R (cadeia lateral) conforme representado na Figura 1.1. A composição do grupo R varia conforme o aminoácido, determinando assim seu tamanho, forma, carga, hidrofobicidade, entre outros aspectos que o tornam os aminoácidos distintos uns dos outros. A Tabela 1.1 apresenta os 20 aminoácidos, juntamente com sua abreviações (3 letras) e seu símbolo (1 letra), como tam-



(a) Estrutura básica de um aminoácido



(b) Glicina, o menor aminoácido



(c) Triptofano, o maior aminoácido

Figura 1.1: Estrutura geral de um aminoácido e exemplos. (a) Em vermelho, apresentamos o grupamento carboxila; em azul o grupamento amina; em verde o carbono-α (CA); e em branco um hidrogênio, formando a cadeia principal. Em laranja, representamos a cadeia lateral através do radical R que irá variar de aminomácido para aminoácido. A glicina representada em (b) é um aminoácido cuja cadeia lateral é composta apenas por um hidrogênio. O triptofano é ilustrado em (c) com sua volumosa e hidrofóbica cadeia lateral.

bém os átomos que formam a cadeia lateral (excluindo os hidrogênios) de seus respectivos resíduos de aminoácidos.

No processo de síntese proteica ocorre formação das ligações peptídicas. Nesse processo, o grupamento carboxila perde uma hidroxila ($-OH$) e ao mesmo tempo o grupamento amina do aminoácido seguinte perde um hidrogênio, deixando ambos com uma ligação livre. Com isso, os aminoácidos se unem, através do OH do grupo carboxila que se liga ao hidrogênio da amina do vizinho e uma molécula de água é liberada no processo, conforme

Aminoácido	Abreviação	Símbolo	Átomos da cadeia lateral
Alanina	ALA	A	CB
Arginina	ARG	R	CB, CD, CG, CZ, NE, NH1, NH2
Asparagina	ASN	N	CB, CG, ND2, OD1
Aspartato	ASP	D	CB, CG, OD1, OD2
Cisteína	CYS	C	CB, SG
Fenilalanina	PHE	F	CB, CD1, CD2, CE1, CE2, CG, CZ
Glicina	GLY	G	*
Glutamato	GLU	E	CB, CD, CG, OE1, OE2
Glutamina	GLN	Q	CB, CD, CG, NE2, OE1
Histidina	HIS	H	CB, CD2, CE1, CG, ND1, NE2
Isoleucina	ILE	I	CB, CD1, CG1, CG2
Leucina	LEU	L	CB, CD1, CD2, CG
Lisina	LYS	Y	CB, CD, CE, CG, NZ
Metionina	MET	M	CB, CE, CG, SD
Prolina	PRO	P	CB, CD, CG
Serina	SER	S	CB, OG
Tirosina	TYR	Y	CB, CD1, CD2, CE1, CE2, CG, CZ, OH
Treonina	THR	T	CB, CG2, OG1
Triptofano	TRP	W	CB, CD1, CD2, CE2, CE3, CG, CH2, CZ2, CZ3, NE1
Valina	VAL	V	CB, CG1, CG2

Tabela 1.1: 20 tipos de aminoácidos comumente encontrados nos seres vivos e seus átomos constituintes. Os átomos da cadeia lateral são designados como β (B), δ (D), γ (G), ϵ (E), e assim por diante, excluindo-se os hidrogênios. *Glicina possui apenas um hidrogênio na cadeia lateral (o que não é considerado nesta abordagem).

Figura 1.2

1.1.2 Estrutura de proteínas

Para entender como uma proteína desempenha a sua função biológica, é essencial saber a sua estrutura tridimensional. Sua conformação nativa se da através do processo de enovelamento, no qual a proteína se dobra em torno de si mesma adquirindo uma estrutura que é determinada pela sequência de aminoácidos no polímero proteico [Stryer et al., 2004]. Em 1934, Bernal & Crowfoot [1934] mostraram que as proteínas, quando cristalizadas, e difratadas por raios-X produziam um padrão complexo de pontos. Apesar de saberem que esses padrões continham as informações necessárias para determinar a estrutura de uma proteína, ainda não era possível para época decifrar esses dados. Em 1958, Kendrew et al. [1958] usou uma técnica aplicada de Max Perutz, que desenvolveu um método para comparar os padrões de cristais contendo diferentes átomos de metais pesados, para produzir as primeiras

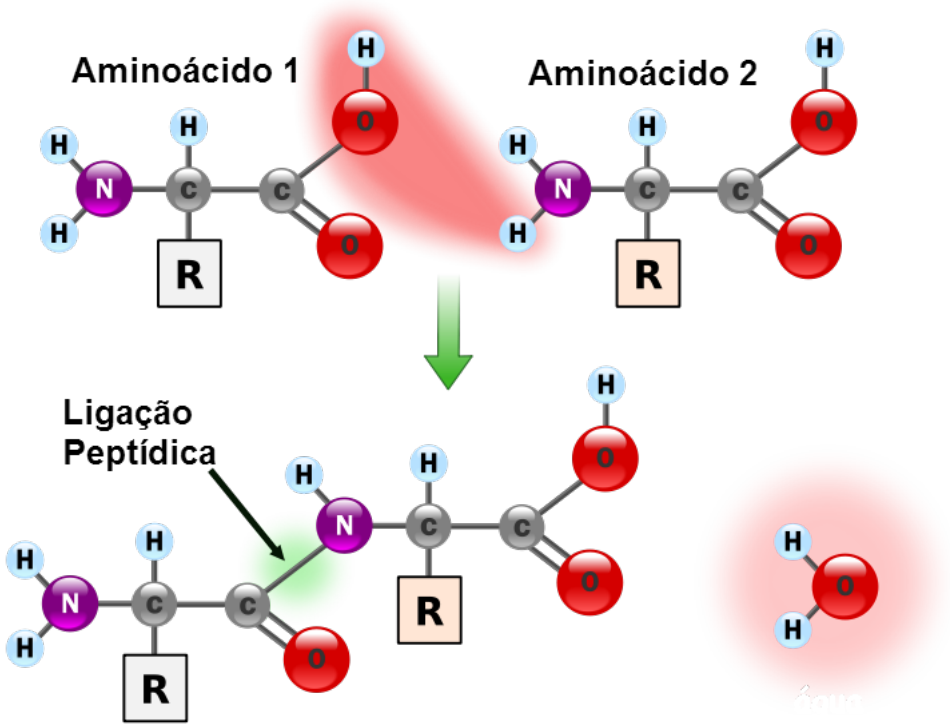
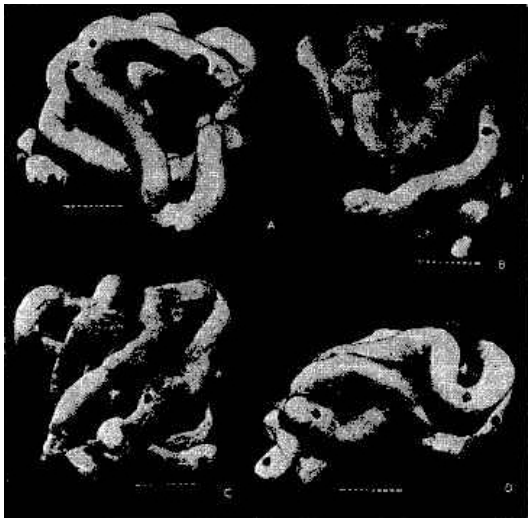


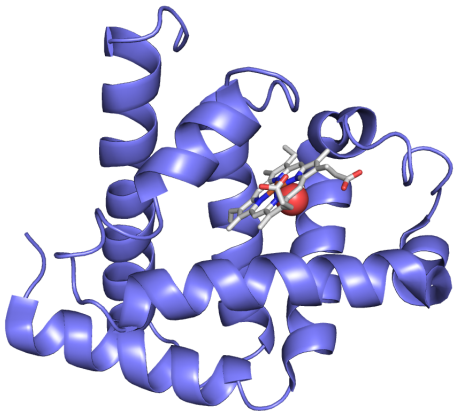
Figura 1.2: Formação de ligação peptídica.

imagens tridimensionais de um proteína - a mioglobina, que é utiliza pelos músculos para armazenar o oxigênio. A Figura 1.3 mostra um comparativo da imagem da mioglobina adquirida a quase 60 anos atrás com as dos dia atuais. Muito do que sabemos atualmente sobre estruturas de proteínas foram graças a esses acontecimentos.

A estrutura de uma proteína tem uma organização que pode ser associada a uma hierarquia. O primeiro nível, chamado de *estrutura primária*, é formado pela sequência linear dos aminoácidos correspondentes a cadeia da proteína. A primeira determinação completa de uma sequência de aminoácidos foi realizada em 1955, por Frederick Sanger, sendo a cadeia B da insulina, formada por 55 resíduos, a primeira sequência de aminoácidos registrada [Sanger, 1988]. As sequências de aminoácidos tendem se enovelar, formando estruturas padronizadas e frequentemente encontradas nas proteínas tais como α-hélices, folhas-β e *loops*, o que se denomina *estrutura secundária* (Figura 1.4). Essas estruturas secundárias ocupam um espaço local na formação das proteínas e com exceção dos *loops*, tendem a seguir uma determinada direção ao longo da cadeia proteica. Estruturas secundárias, por sua vez, também se conectam fazendo com que a proteína assuma um forma tridimensional, denominada *estrutura terciária*. Neste nível podemos dizer que a proteína é estabilizada, havendo várias forças atuando sobre os resíduos e átomos que a compõem, mantendo sua



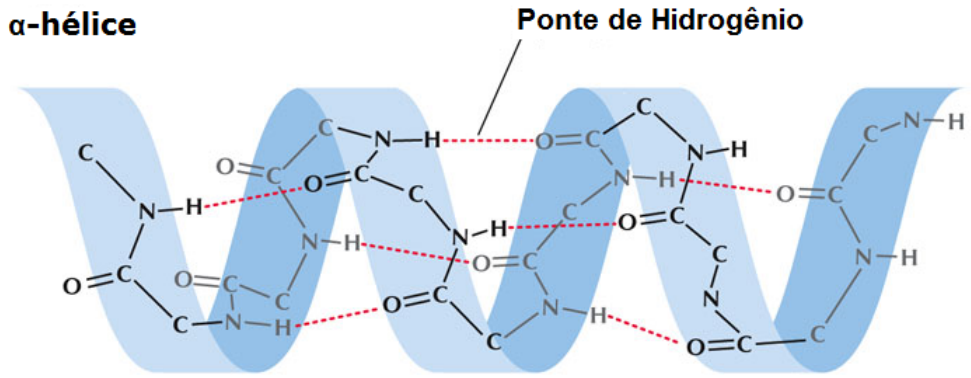
(a) Primeira imagem de um estrutura de protéina (mioglobina) [Kendrew et al., 1958].



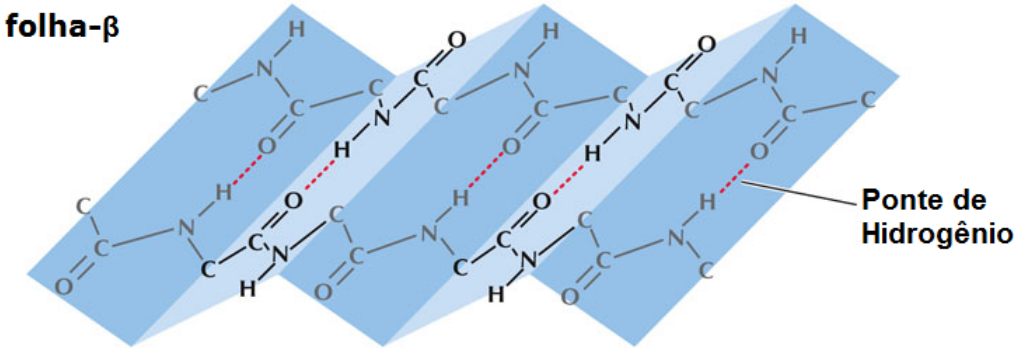
(b) Imagem da mioglobina gerada nos dias atuais. Imagem criada a partir do programa Pymol [Delano, 2002].

Figura 1.3: O primeiro registro de uma imagem de uma proteína em contraste com os modelos atuais.

conformação estrutural. Algumas proteínas possuem mais de uma cadeia, que em geral são enoveladas separadamente como estruturas terciárias e depois se unem a outras para formar um complexo biologicamente ativo. Esse tipo de complexo se refere ao último nível da hierarquia, denominado *estrutura quartenária*. O presente trabalho analisa dados referentes aos contatos estabelecidos entre as cadeias destes complexos quaternários.



(a) α-Hélices.



(b) Folhas-β.

Figura 1.4: Estruturas secundárias. Adaptada de: [Geoffrey M. Cooper, 2006]

1.2 Interações em Proteínas

Conhecer e compreender como as interações covalentes e não-covalentes entre proteínas ocorrem nas células permitiria descrever com abrangência os vários mecanismos e funções celulares [Franceschini et al., 2013]. Interações não-covalentes são muito comuns em proteínas, devido ao fato de que estas são moderadamente mais fracas, comparadas as interações covalentes, que requerem bem mais energia para serem quebradas [Kessel & Ben-Tal, 2010]. Além disso, interações não-covalentes permitem com que as proteínas mudem sua conformação, dando flexibilidade a certas regiões e possibilitando a interação com ligantes, duas propriedades de suma importância para manter suas funções. Embora estas ligações sejam fracas, em grande quantidade elas permitem com que as proteínas se mantenham estáveis quanto a sua estrutura tridimensional nativa. É importante ressaltar que a natureza e força das interações não-covantes em proteínas são afetadas pelo ambiente ao qual as mesmas estão inseridas [Kessel & Ben-Tal, 2010]. Ainda quanto as interações não-covalentes,

pode-se classificá-las em três grupos:

- Interações eletrostáticas: entre átomos carregados eletricamente, onde se incluem as pontes de hidrogênio, pontes salinas e empilhamentos aromáticos.
- Interações de van der Waals: entre pares de átomos quaisquer quando estão muito próximos um do outro, criando forças de atração e repulsão.
- Interações apolares: resultantes de efeito hidrofóbico que se observa com maior frequência entre átomos ou grupos químicos neutros e hidrofóbicos.

1.2.1 Ligações de hidrogênio

Pontes de hidrogênio ou ligações de hidrogênio são interações de caráter eletrostático que ocorrem entre átomos com diferentes eletronegatividades. Muito comuns tanto em proteínas como entre estruturas de DNA, estas interações desempenham papéis importantes como, por exemplo, conferir a especificidade da conformação da proteína, como em estruturas secundárias ou associando com outras moléculas; ajudando no enovelamento da proteína; auxiliando processo de catálise enzimática [Fogolari et al., 2002]. Ligações de hidrogênio ocorrem quando um átomo de hidrogênio ligado covalentemente à um átomo denominado "doador"interage com um átomo eletronegativo como, por exemplo, o nitrogênio (N), o oxigênio (O) ou o flúor (F), denominado "aceptor"e sendo suas forças de interação dependentes do alinhamento dos átomos envolvidos [de Melo et al., 2007]. Além disso, são responsáveis pelas propriedades que a água possui quanto à sua capacidade de solvente universal. [Kessel & Ben-Tal, 2010].

1.2.2 Interações hidrofóbicas

As interações hidrofóbicas ocorrem entre grupos não polares. A relativa ausência de interações entre moléculas apolares e a água provoca interações entre átomos desse grupo que parecem ser mais favoráveis do que seriam em outros solventes. Dessa forma as moléculas apolares preferem ambientes do mesmo tipo. Esta preferência por ambientes não carregados é conhecida como efeito hidrofóbico.

Em proteínas globulares, o efeito hidrofóbico é importante por manter os átomos num arranjo tal que átomos com maior polaridade ou hidrofílicos permaneçam na superfície externa da proteína, podendo interagir com outras moléculas, e átomos hidrofóbicos tendam a permanecer no interior da proteína [Mancini et al., 2004]. Embora as interações hidrofóbicas sejam relativamente fracas comparadas às pontes de hidrogênio e pontes salinas, elas acabam tendo grande relevância no processo de enovelamento de proteínas, sendo a principal

força atuante neste processo e desempenhando um papel dominante nas interações proteína-proteína [Tsai et al., 1997].

1.2.3 Pontes salinas

Pontes salinas são interações eletrostáticas entre átomos com carga formal. Ocorrem entre ânions (átomos carregados negativamente) e cátions (átomos carregados positivamente). Uma forma de descrever este tipo de interações seria utilizando as propriedades da equação de Coulomb, para definir que átomos com cargas de mesmo sinal se repelem (interação repulsiva) e cargas com sinais diferentes se atraem (interação atrativa). Esse cálculo pode ser inviável em larga escala pelo custo computacional. Uma abordagem mais simples consiste em considerar somente cargas formais na proteína, ou seja, onde um elétron foi doado ou recebido [Bickerton et al., 2011]. Há ainda métodos que consideram os valores de pK_a para resíduos carregados, mas também não são viáveis para processos de larga escala [Davies et al., 2006].

1.2.4 Empilhamentos aromáticos

Pontes salinas e pontes de hidrogênio representam a maior parte das interações eletrostáticas em proteínas, porém há alguns grupos químicos capazes de participar destas interações não-covalentes. Entre eles estão os anéis aromáticos que são encontrados nos aminoácidos de HIS, PHE, TRP e TYR, e que podem estabelecer empilhamentos aromáticos. Os anéis aromáticos possuem ligações duplas em ressonância, adquirida por conta dos movimentos cíclicos que os elétrons realizam nos orbitais. Elétrons no orbital σ (dentro do plano do anel) fazem com que se criem cargas parcialmente positivas no plano do anel e os elétrons no orbitais π, localizados na parte superior e inferior do plano do anel desenvolvem uma carga parcialmente negativa nestas regiões. Devido à existência destas cargas parciais, anéis aromáticos podem interagir uns com os outros o que se denomina empilhamento aromático [Kessel & Ben-Tal, 2010].

1.3 Definição de contatos em proteína

Atualmente a literatura apresenta várias abordagens para definição de contatos em proteínas [Silveira et al., 2009]. Segundo Mancini et al. [2004], contatos inter-atômicos são forças de atração ou de repulsão existentes entre átomos distintos. É importante ressaltar que em alguns casos os termos para contato e interação pode ser distinto. Silveira et al. [2009] descreve contato como um termo que se refere apenas à posição e distribuição dos átomos

espacialmente, utilizada para denominar apenas a vizinhança do átomo ou resíduo. O termo de interação, em contrapartida, se refere às forças mútuas exercidas entre os átomos, como atração e repulsão devido suas polaridades, por exemplo.

Contatos podem ser analisados tanto a nível de resíduos [Miyazawa & Jernigan, 1985] como a nível de átomos [Sobolev et al., 1999; Mancini et al., 2004]. A nível atômico, é possível realizar uma análise mais refinada e detalhada quanto as propriedades das interações em comparação a quando se faz estudos à nível residual. Porém, há uma ordem de grandeza maior de processamento e informação para lidar quando se utiliza nível atômico.

Determinar com exatidão e precisão os parâmetros para cálculo de contatos é de grande importância para utilização de algoritmos que façam análises ou comparações em estruturas de proteínas e há uma grande quantidade estudos sobre isso, como por exemplo, em alinhamento estrutural [Holm & Sander, 1993], predição de estrutura [Samudrala & Moult, 1998; Bowie et al., 1991], interação proteína-proteína [Bickerton et al., 2011] e proteína-ligante [Fassio, 2015].

Um dos métodos mais clássicos e simples para definir contatos consiste em estabelecer delimitadores de distância para os diversos tipos de interação. Dado um par de pontos no espaço *i,j*, sendo esses pontos a posição de um átomo ou resíduo, *i* estará em contato com *j* se a distância entre esses dois pontos satisfaz um critério definido. Vasculhando a literatura, percebe-se que há um problema na definição de distância, pois existem muitas opções a se usar. Em nível atômico verifica-se distâncias tais como 3,8 Å [Mancini et al., 2004], 5,0 Å [Godzik et al., 1992], 6,0 Å [Plaxco et al., 1998], e em alguns casos esta distância varia conforme as propriedades físico-químicas entre os átomos envolvidos [Bickerton et al., 2011]. Já em nível de resíduo encontra-se 6,5 Å [Miyazawa & Jernigan, 1985], 7,0 Å [Silveira et al., 2009], 8,0 Å [Manavalan & Ponnuswamy, 1977]. Portanto, não há consenso sobre o valor mais adequado para esse delimitador de distâncias, sendo estes muitas vezes escolhidos arbitrariamente ou para atender a alguma otimização necessária.

Então, para saber se dois átomos estão em contato é preciso definir como calcular ou qual método utilizar antes do início do um processo que venha a utilizar contatos ou interações em proteínas.

Segundo Silveira et al. [2009] as duas abordagens relevantes para cálculos de contato são usando: delimitadores dependentes de distância de corte (*cutoff dependent*) e delimitadores independentes (*cutoff free*). O primeiro cálculo define um contato entre pares de resíduos ou átomos (dependendo do nível utilizado) se a distância Euclidiana entre o seus centros (no caso de resíduos, calcula-se o centroide de cada um) forem menor ou igual ao um valor estabelecido. Já para contatos definidos por delimitador independente pode-se usar um método de cálculo geométrico conhecido como Triangulação de Delaunay.

1.3.1 Triangulação de Delaunay

O uso da técnica de Triangulação de Delaunay em proteínas remonta a Richards [1974], usados para cálculo de volume e densidade, e continua sendo utilizado em trabalhos recentes [Silveira et al., 2009; Fassio, 2015]. O método dual da Triangulação de Delaunay é o Diagrama de Voronoi. Ele usa uma forma de cobertura de espaço, seja ele qualquer dimensão, preenchendo-o de forma justa com relação a cada ponto p utilizado no conjunto. Com isso, estes métodos são capazes de capturar relações espacias entre o conjunto de pontos p. A Triangulação de Delaunay e o Diagrama de Voronoi, através de regras geométricas exatas, produzem um tipo de conectividade envolvendo sempre os vizinhos mais próximos de cada ponto p. Em proteínas, o uso da Triangulação de Delaunay resulta em um decomposição do volume ocupado pelos átomos em tetraedros justapostos, organizados de tal forma que os contatos representarão arestas e os átomos ou resíduos serão vértices.

A Figura 1.5 ilustra o uso da Triangulação de Delaunay para um espaço tridimensional, o qual é o mesmo quando lidamos com estrutura de proteínas. Observa-se que as triangulações são substituídas por tetraedros.

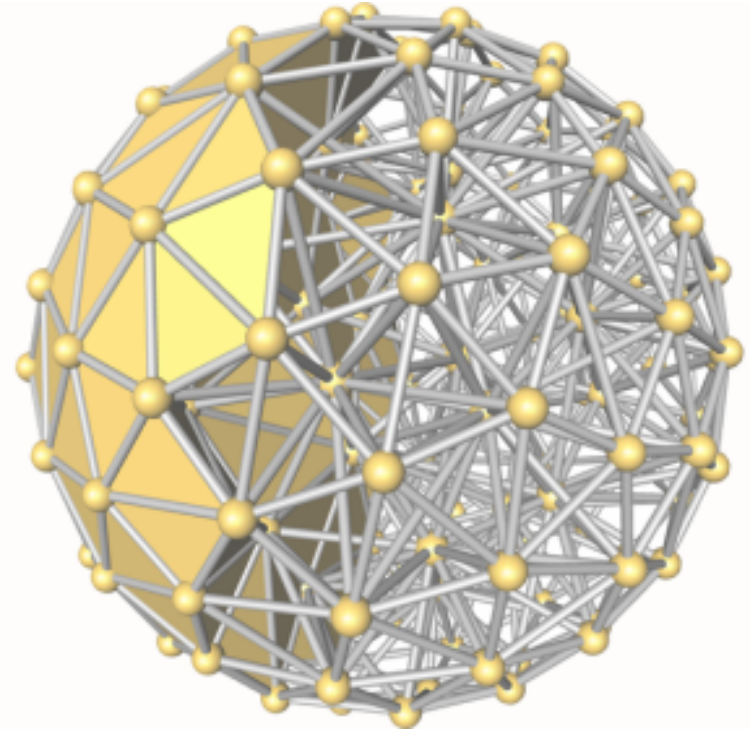


Figura 1.5: Exemplo de Triangulação de Delaunay no espaço tridimensional. Fonte: http:doc.cgal.orglatestTriangulation_3

A vantagem de uso desta técnica em contatos de proteínas se deve ao fato de que ela organiza as arestas entre os vértices de forma justa, evitando a ocorrência de sobreposições, falhas ou buracos [Silveira et al., 2009].

1.3.2 Delimitador Dependente

Mais simples de ser implementado que a Triangulação de Delaunay, o método para cálculo de contato em proteínas utilizando delimitadores dependentes em uma distância de corte específica. Necessita três argumentos para ser calculado: o centro espacial primeiro átomo (ou centroide do primeiro resíduo), o centro espacial do segundo e a distância de corte definida, sendo que a omissão ou a atribuição de um valor elevado deste último parâmetro pode ocasionar em um cálculo de contatos entre todos os pares de átomos ou resíduo da proteína, resultando em uma avaliação de todos-contra-todos ao qual podem gerar muito contatos falso-positivos Para calcular a distância entre os centros dos pares, usa-se a equação de distância euclidiana, onde quando se usa três dimensão é defina conforme equação:

$$D(i,j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

Ao qual, $D(i,j)$ é a distância entre os átomos i e j, e suas coordenadas no espaço tridimensional são definidas por (x_i, y_i, z_i) e (x_j, y_i, z_i), respectivamente. Conforme abordado na Seção 1.3 não há um distância de corte fixa definido na literatura e a sua escolha varia conforme adequação as condições.

Para ilustrar a diferença entre métodos de Delimitador Dependente e por Triangulação de Delaunay tomemos como exemplo a Figura 1.6.

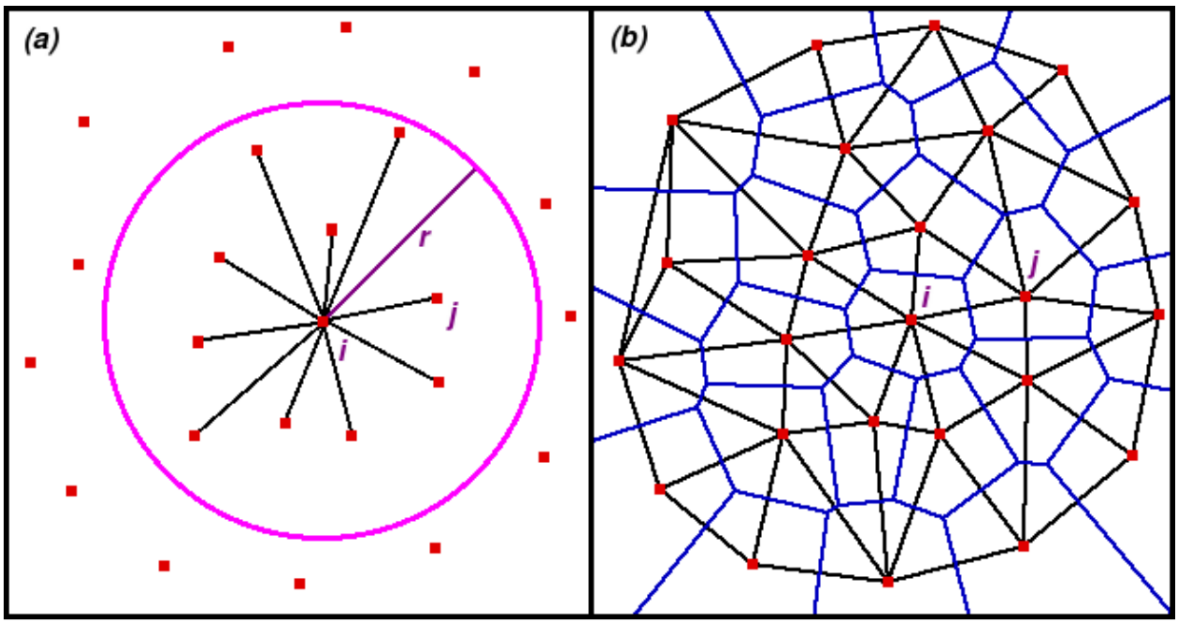


Figura 1.6: Exemplo do método de Delimitador Depentente (a) e a Triangulação de Delaunay (b) em um espaço bidimensional [Silveira et al., 2009].

A Figura 1.6(a) demostra o método de Delimitador Dependente onde os pontos (qua-

drados vermelhos) representando os vértices e o círculo (roxo) a vizinhança do vértice central *i*, dentro do delimitador *r*. A Figura 1.6(b) apresenta a aplicação da Triangulação de Delaunay juntamente com seu dual, Diagrama de Voronoi nos mesmos conjuntos pontos de (a) sem um delimitador *r*. Pode-se observar as células (arestas em azuis) criadas pelo Diagrama de Voronoi sem o delimitador de distância (r) faz com que a Triangulação de Delaunay (arestas em preto) criam uma malha ao qual os pontos (quadrados vermelhos) só se conectam com seus vizinhos mais próximos.

1.4 Banco de Dados Biológicos

Recentemente estamos presenciando um crescimento exponencial da quantidade de dados biológicos, e estima-se que esses dados continuaram crescendo nos próximos anos. Um passo natural para lidar com este tipo de acontecimento é fazer o uso do conhecimento computacional para organizar e gerir esta demanda de dados, combinando a biologia com a computação, e com isso dando origem à disciplina da bioinformática. Com vários projetos científicos sendo executados ao redor do mundo, o volume de informação de entrada é gigantesco e é necessário organizar e arquivar estes resultados. O arquivamento e disponibilização desses dados são realizados por organizações que mantêm imensos banco de dados de diversas especificidades [Lesk & Andrade, 2008].

Outrora, o arquivamento de dado na área da bioinformática eram mantidos por grupos de pesquisa individuais. Com o aumento da demandas por pessoas e equipamentos este tipo de arquivamento passou a ser responsabilidade de projetos nacionais e internacionais em uma grande escala. Muito projetos, inicialmente modestos e com objetivos simples, ascenderam ao ponto de ser de extremo interesse da industrial multinacional, com a compra e venda de empresas do ramo. Abaixo, segue a lista de alguns banco de dados primários de macromoléculas biológicas:

- Sequência de ácidos nucléicos e genomas completos [Leinonen et al., 2010].
- Sequência de aminoácidos de proteínas [Boeckmann et al., 2003].
- Estrutura de proteína e ácidos nucléicos [Berman et al., 2000].
- Classificação estrutural de protéinas [Murzin et al., 1995]

1.4.1 Protein Data Bank - PDB

O Protein Data Bank (PDB) é o principal repositório internacional de arquivos que contêm informações sobre as estruturas tridimensionais de macromoléculas biológicas, in-

cluindo proteínas e ácidos nucleicos. Estas macromoléculas depositas provêm dos mais diversos tipos de organismos vivos, incluindo bactérias, leveduras, plantas entre outros animais e seres humanos [Berman et al., 2000]. Foi iniciado por Walter Hamilton do Brookhaven National Laboratories, Long Island, Nova York, em 1971 e é gerenciado atualmente pelo Research Collaboratory for Structural Bioinformatics (RCSB), a qual esta distribuída por vários estados dos Estados Unidos.

O endereço eletrônico[1] do PDB possui uma interface para buscar os próprios arquivos de dados, além de recursos para depósito de novas entradas e *softwares* especializado para recuperação e análises de estruturas.

Em 2003, o RCSB, o Molecular Structure Database, o European Bioinformatics e o Protein Data Bank of Japan se uniram e formaram o Worldwide Protein Data Bank (wwPDB), com o intuito de produzir um formato unificado de arquivo.

Os arquivos PDB são os principais documentos que contêm informações sobre estruturas de proteínas, onde se incluem:

- O nome da proteína e o assunto da entrada, juntamente com a espécie ao qual ela pertence.
- Os autores que determinaram a estrutura e as referências para publicações.
- Detalhes experimentais, tais como qualidade geral do resultado e qual método utilizado.
- A sequência primária da estrutura.
- Moléculas adicionais, como inibidores e solventes.
- Informações sobre estruturas secundárias.
- Coordenadas atômicos.

Desde sua criação em 1971 com apenas sete estruturas depositadas, a quantidade de entradas de arquivos vem crescendo exponencialmente, conforme se observa na Figura 1.7, devido a avanços em resolução de estrutura de proteínas com ressonância magnética Atualmente conta com mais de 110.000 estruturas de proteínas conhecidas, sendo sua grande maioria resolvida por cristalografia utilizando difração de raio X ou por ressonância nuclear magnética.

[1] http://www.rcsb.org/

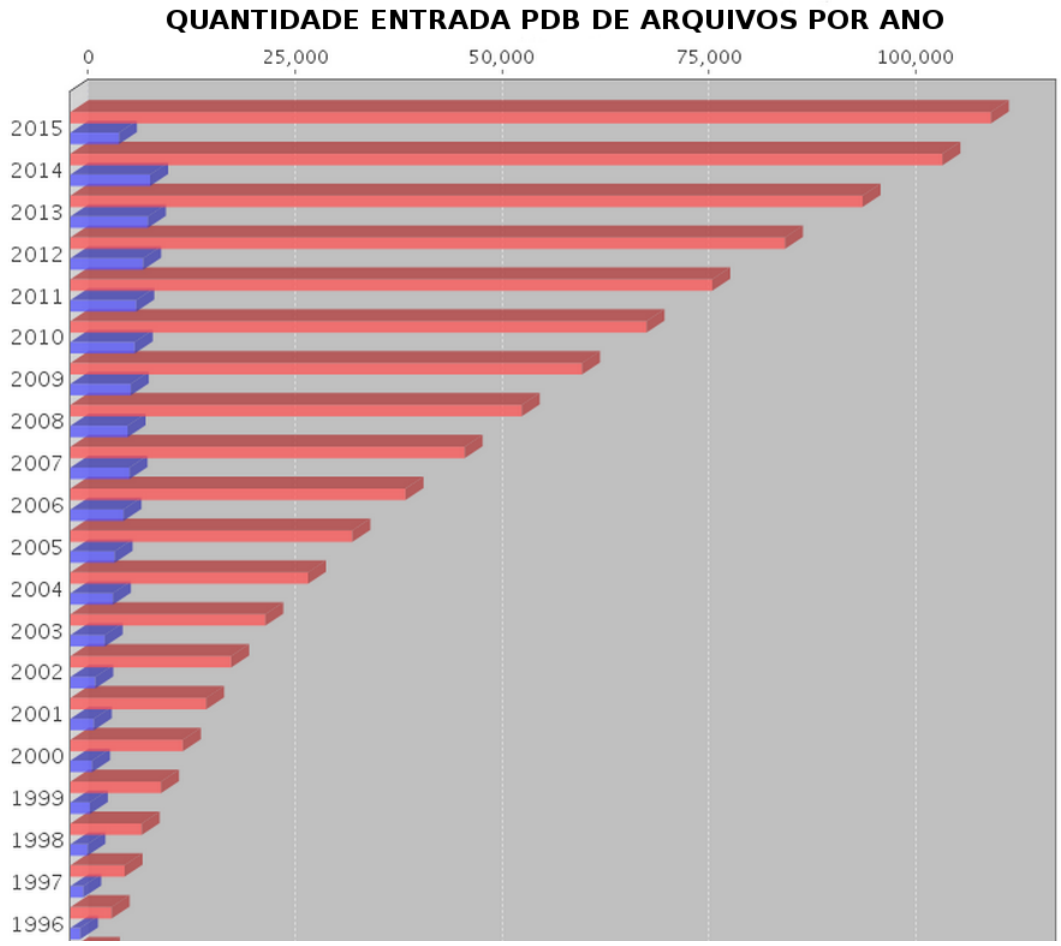


Figura 1.7: Gráfico de barras da quantidade anual (em azul) de arquivos PDB. As barras vermelhas representam o montante total. Imagem adaptada de: www.rcsb.orgpdbstatistics

1.4.2 PICCOLO

O PICCOLO é um banco de dados relacional de interações proteína-proteína desenvolvido por Bickerton et al. [2011]. As interações são descritas no nível atômico, apresentando interações de pares de átomos, resíduos e cadeias de um complexo polipeptídico, utilizando critérios de distância e termos de ângulo para definir 12 tipos diferentes de interação, entre elas: van der Waals, ponte de hidrogênio, interações hidrofóbicas, pontes salinas. Para o seu desenvolvimento foi necessários utilizar programas para calcular a acessibilidade ao solvente [Hubbard & Thornton, 1993], para definir os resíduos da interface, e também de calculo de ponte de hidrogênio [McDonald et al., 1993].

Inicialmente continha 38,202 complexos proteico, e hoje já dispõe de 46.805 complexos, totalizando cerca de 230 milhões de interações de pares de átomos.

Apesar das informação contidas no PICCOLO serem estáticas, uma vez que não houve

mais atualização[2] quanto a novas estruturas comparado ao que se vê hoje do PDB, as estruturas contidas no PICCOLO são bem acuradas e foi de grande inspiração e utilidade para o presente trabalho.

1.5 Objetivos

1.5.1 Objetivos Gerais

Comparar diferentes paradigmas para cálculo de contatos em interfaces proteína-proteína, afim de analisar e estimar critérios de distância entre interações atômicas e tentar identificar o paradigma mais acurado.

1.5.2 Objetivos Específicos

- Estudar os diferentes paradigmas que são usados para cálculo de contatos em interfaces proteína-proteína.
- Definir um conjunto de arquivo de estruturas de proteína que serão usadas como fonte de dados para o estudo comparativo.
- Computar os contatos segundo os diferentes paradigmas através do desenvolvimento de um programa que seja capaz de lidar com dados em larga escala.
- Projetar e implementar um banco de dados que armazenasse as interações proteína-proteína dos paradigmas estudados.
- Projetar e implementar visualizações interativas que possibilitassem a exploração, observação e análise dos resultado obtidos com o intuito de comparar os diferentes paradigmas.
- Disponibilizar os artefatos desenvolvidos neste trabalho para a comunidade científica.

[2]Última versão datada em: 15 de fevereiro de 2013

Capítulo 2

Materiais e Métodos

Neste capítulo, apresentamos os métodos e as ferramentas utilizados para construção do banco de dados CAPRI e a aplicação web que será responsável pela apresentação e análise dos resultados obtidos com as comparações dos paradigmas para prospecção de contatos estudados.

2.1 Sistema gerenciador de bancos de dados e linguagens de programação utilizados

A base de dados foi desenvolvida utilizando o sistema gerenciador de banco de dados (SGBD) MySQL. Esta escolha se deve não somente ao fato de ser o SGBD utilizado pela base de dados PICCOLO, mas principalmente por ser uma ferramenta gratuita e robusta, já amplamente utilizada pela comunidade acadêmica. Para povoar o nosso banco de dados, utilizamos a linguagem de programação Python que é simples e de fácil manutenção, além de possuir bibliotecas que auxiliaram muito no objetivo do trabalho [Hamelryck & Manderick, 2003; Jones et al., 2001].

Para criação das paginas da aplicação web foi utilizada a linguagem de programação interpretada *JavaScript*, com suas bibliotecas JQuery[1], para manipulação de dados de forma dinâmica e D3.js[2] [Bostock et al., 2011] para criação de gráficos interativos. Além disso, o desenho das páginas foi feito através do *framework* Bootstrap[3] que facilita a organização do *layout* da aplicação.

[1] http://jquery.com/
[2] http://d3js.org/
[3] http://www.getbootstrap.com

2.2 Paradigmas para prospecção de contatos

Para computar contatos, existem quatro fatores principais ou critérios que podem ser levados em consideração:

Propriedades fisico-químicas: são essenciais na definição dos tipos de interações que ocorrem entre os pares de átomos envolvidos, por exemplo, estando dois átomos hidrofóbicos em uma certa distância eles podem fazer uma interação hidrofóbica. A lista das propriedades fisico-químicas para cada átomo está disponível no Apêndice A.

Distância: distância euclidiana entre pares de pontos que representam o centro dos pares de átomos envolvidos no contato. Essas distâncias comumente são valores tabelados e definidos de forma bastante *ad hoc* nos trabalhos que utilizam a distância para definição de contatos.

Oclusão: a utilização apenas do critério de distância entre pares de átomos na definição de um contato leva frequentemente à indicação de contatos que são falsos positivos devido à existência de outro(s) átomo(s) ocluindo a interação e fazendo com que a mesma não seja legítima. Para contornar esse problema, utilizamos uma abordagem geométrica baseada no método de Diagrama de Voronoi, mais especificamente o seu dual chamado Triangulação de Delaunay. Esse método constrói geometricamente poliedros ligando pontos representativos dos átomos e garante que pares de átomos oclusos por um terceiro átomo nunca sejam conectados por arestas. A Figura 2.1 exemplifica este fenômeno: existe um nitrogênio do resíduo 217 da cadeia B (azul) realizando ponte salina com três outros átomos da cadeia A: um oxigênio do resíduo 107 (verde); e um carbono e um oxigênio do resíduo 110 (amarelo). O átomo de carbono do resíduo 217 (vermelho), utilizando o método de *delaunay*, impede a formação de uma aresta (reta tracejada vermelha) entre os átomos do resíduo 110, criando somente um contato entre oxigênio do resíduo 107 e o nitrogênio do resíduo 217 (reta tracejada verde).

Ângulos: ligações de hidrogênio e empilhamentos aromáticos são interações cuja força de interação sabidamente depende do ângulo formado entres os grupos interagentes. Por esse motivo, acreditamos que métodos que levam em conta a angulação entre os grupos tendem a ser mais precisos na propospecção deste tipo de contato.

Neste estudo, comparamos três diferentes paradigmas para prospecção de contatos baseados em distância, oclusão e distância, distância e ângulação, sendo todos eles dependentes das mesmas definições de propriedades fisico-químicas. Esses paradigmas são melhor explicados na Seção 2.2.1.

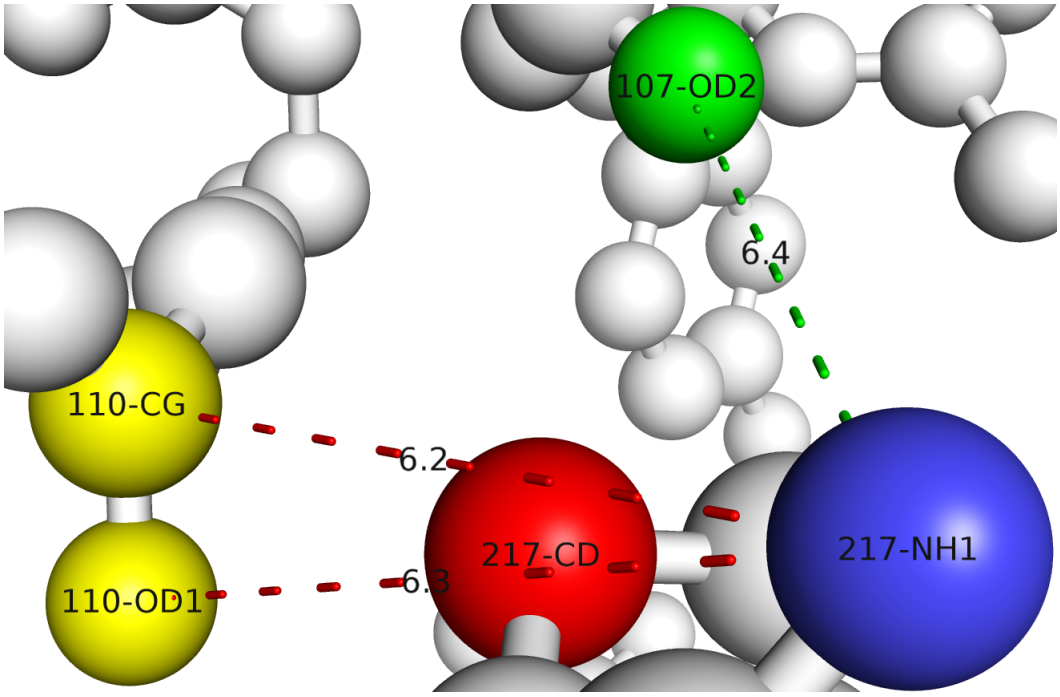


Figura 2.1: Exemplo de oclusão pela Triangulação de Delaunay. Interface do arquivo PDB 1SNE. Arestas em vermelho mostram os contatos oclusos pelo átomo 217-CD enquanto a verde ilustra um contato genuíno.

2.2.1 Paradigmas comparados

Os paradigmas usados na prospecção de contatos que foram comparados neste estudo são os seguintes:

1. **Delimitador dependente (*cutoff*):** Utiliza raios de varredura fixados em até 7.0 Å para todos os contatos. Não evita oclusões e não usa critérios de angulação. É o método mais simples de ser implementado.

2. **Triangulação de Delaunay (*delaunay*):** Utiliza raios de varredura fixados até 7.0 Å para todos os contatos. Evita oclusões pois a triangulação garante que apenas a primeira camada de átomos vizinhos são conectados por arestas. Não utiliza critérios de angulação.

3. **Piccolo:** Utiliza raios de cortes distintos (conforme Tabela 2.1) para cada tipo de contato. Não evita oclusões e usa critérios de angulação para ligações de hidrogênios e empilhamentos aromáticas.

Quando se usa critérios de angulação para definir ligações de hidrogênio é necessário saber a localização espacial do átomo de hidrogênio ligado ao átomo doador, pois o ângulo formado entre átomo doador, o átomo de hidrogênio e o átomo aceptor determina a força

Tipo de contato	Critério de Distância	Critério de Angulação
Ligação de hidrogênio	d(a_i, a_j) <3,9 d(a_h, a_a) <2.5	θ(a_d,a_h,a_a) >90 ° θ(a_d,a_a,a_{a-ant}) >90 ° e θ(a_h,a_a,a_{a-ant}) >90 °
Interação hidrofóbica	d(a_i, a_j) <5	-
Ponte salina	d(a_i, a_j) <6	-
Empilhamento aromático	d(a_i, a_j) <6	-

Tabela 2.1: Contato entre átomos *i* e *j* baseado em suas propriedades, sendo (d) a distância e θ o critério de angulação. θ(a_1,a_2,a_{3t}) representa o ângulo em a_2 entre a_1 e a_3; a_d = átomo doador; a_a = átomo aceptor; a_h = átomo de hidrogênio do doador; a_{a-ant} = átomo antecedente ao átomo aceptor.

desta interação, sendo proporcional ao ângulo. No caso do PICCOLO foi usado o programa HBPLUS [McDonald et al., 1993] que adiciona átomos de hidrogênios à proteína e depois detecta as ligações de hidrogênio considerando a distância dos átomos aceptores e doadores de até 3,9 Å e a angulação dos átomos envolvidos (aceptor, hidrogênio e doador) maior que 90°.

É importante ressaltar que, no intuito de comparar os três paradigmas, tivemos que trabalhar no mesmo escopo para os dados computados pelos diferentes métodos. Dessa forma, destacamos que os três métodos avaliam contatos em interfaces de proteína, ou seja são contatos inter-cadeias. Adicionalmente, as listas dos resíduos e átomos que consideramos como válidos neste trabalho são o mesmos apresentados na Introdução, na Tabela 1.1. Os tipos de contatos analisados são:

- Ligação de hidrogênio
- Interação hidrofóbica
- Pontes salinas (atrativa)
- Empilhamento aromático.

2.3 Modelagem da base de dados

Para cada paradigma estudado criamos uma tabela de mesma estrutura tendo o prefixo denominado como *atom_pairs_X*, onde *X* seria a variação do nome adequado ao paradigma referente, sendo assim denominadas:

- atom_pairs_cutoff.
- atom_pairs_delaunay.

- atom_pairs_piccolo.

A tabela que contém as informação dos pares de contatos na base de dados PICCOLO foi adaptada para se comportar ao devido propósito deste trabalho. Sendo assim alguns campos foram alterados e removidos. A única remoção de registros realizada nesta tabela foi quanto aos identificadores PDB que não puderam ser comparados por motivos de obsolescência ou ocorrência de revisão de arquivo, desde a última atualização realizada pela base de dados PICCOLO[4]. A seguir, apresenta-se os campos que são comuns entre as três tabelas:

- **pdb**: identificador PDB referente ao arquivo.
- **p1_chain**: identificador da cadeia do primeiro átomo.
- **p2_chain**: identificador da cadeia do segundo átomo.
- **p1_resid**: identificador do resíduo do primeiro átomo.
- **p2_resid**: identificador do resíduo do segundo átomo.
- **p1_resname**: sigla do resíduo do primeiro átomo.
- **p2_resname**: sigla do resíduo do segundo átomo.
- **p1_atname**: sigla do primeiro átomo (Conforme Tabela 1.1).
- **p2_atname**: sigla do segundo átomo (Conforme Tabela 1.1).
- **is_contact**: etiqueta que indica se houve um contato.
- **is_hb**: etiqueta que indica se houve um contato de ponte de hidrogênio.
- **is_hydrophobe**: etiqueta que indica se houve um contato de hidrofóbica.
- **is_ionic**: etiqueta que indica se houve um contato de ponte salina (atrativa).
- **is_aromarom**: etiqueta que indica se houve um contato aromático.
- **is_proximal**: etiqueta que indica se os átomos estão dentro do raio definido.
- **distance**: distância do contato (em Å).

[4]Última atualização em 15 de Fevereiro de 2013. http://www-cryst.bioc.cam.ac.uk/~richard/PICCOLO/downloads.php. Último acesso 15/07/2015

Método experimental	Quantidade	Percentual
Difração de raio X	44.123	97,01 %
RNM	1.082	2,37 %
Outros	275	0,60 %

Tabela 2.2: Frequência dos métodos de resolução de estruturas de proteínas na base de dados utilizada.

Além disso, para controlar o processo de inserção de dados nas tabelas dos paradigmas *cutoff* e *delaunay*, foi criado uma tabela de controle, denominada *loading_log* que contém informações sobre os arquivos a serem carregados na base como quantidade de cadeias, resíduos e átomos, além de armazenar o experimento utilizado para resolução da estrutura e, em caso de difração de raio X, a sua resolução e a data da última atualização do arquivo. A Tabela 2.2 contêm a frequência de cada método de resolução de estruturas para o conjunto de dados usado neste trabalho.

2.4 Carga no banco de dados CAPRI

O nome *CAPRI* provém do acrônimo *Comparative Analysis of Protein-pRotein Interaction* e esta foi a base de dados criada para armazenar todos os dados a serem comparados, bem como para facilitar e agilizar as inúmeras consultas que foram necessárias para as análises feitas. Para a carga e consultas, este projeto exige grande capacidade de processamento e armazenamento. Apenas para ilustrar esse ponto, levou-se cerca de três dias para computar os contatos de cerca de 45.000 arquivos do PDB, referentes as estruturas de protéinas encontradas na base de dados PICCOLO. Utilizamos duas máquinas trabalhando em paralelo, ambas com sistema operacional Linux (distribuição Ubuntu 12.04), processador Intel Core 2 Duo (3Ghz) e 4 GigaBytes de memória RAM. A Tabela 2.3 possui algumas informações sobre a quantidade média e desvios padrão das cadeias, resíduos e átomos por proteína, provindo dos arquivos PDB que constam no nosso estudo.

Por ser um processo extremamente custoso computacionalmente, criamos um controle de carga para permitir que o processo continuasse em caso de eventuais erros ou paradas do programa que viessem a ocorrer. Quando um determinado arquivo PDB sinaliza uma falha, a mesma é reportada, informando o momento e a descrição do problema e o processo continua com o próximo arquivo PDB. O relatório de erros é então analisado e assim, caso os problemas possam ser solucionados com os arquivos que possuem essas questões, os mesmos são marcados com uma etiqueta *"READY"* (pronto) indicando que está pronto para ser processado novamente. Arquivos PDB que têm seu processamento concluído com sucesso

Nº arquivos	Cadeias		Resíduos		Átomos	
	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ
45.480	3,59	3,01	824,59	790,84	6444	6142

Tabela 2.3: Resumo sobre o volume da base de dados utilizada nesse trabalho. A saber: $\bar{X}$: média; σ: desvio padrão.

são marcados como *"OK"*, evitando assim o seu reprocessamento em eventuais problemas que demandem o recomeço do povoamento da base. Esse processo, embora trabalhoso de ser projetado e implementado, foi extremamente útil nos inúmeros eventos que ocorreram durante o desenvolvimento do projeto como quedas de energia, travamentos do sistema operacional ou problemas de hardware.

A Figura 2.2 demostra o diagrama de fluxo utilizado pelo programa no processo de carga. Cada arquivo PDB passa pelo fluxo, e em caso de erro, como já mencionado, é enviado um registro na tabela de controle e o ciclo é interrompido e retorna para a estado inicial de *Filtragem dos Arquivos PDB* com o arquivo PDB seguinte. Arquivos temporários são utilizados para saída e entrada entre cada etapa e são excluídos a cada ciclo, independentemente se este foi bem sucedido ou não. A seguir detalhamos as fases deste processo.

A primeira fase consiste em verificar a lista de arquivos PDB que são utilizados como entrada no processo. Para isso, se fez necessário adquirir todos os arquivos PDB no portal WWPDB[5]. No total, realizamos o *download* de 46.805 arquivos de estruturas, conforme observado na tabela de pares de contatos do banco de dados PICCOLO, porém, 45.480 (representando 97,18%) foram processados com sucesso. Isso ocorreu devido a obsolescência e substituição de arquivos como previsto nas políticas de processamento de documentos do portal[6].

Após a coleta dos arquivos, realizamos os passos a seguir utilizando um programa desenvolvido em Python para processar os arquivos PDB, determinando os pares de átomos da interface, os contatos estabelecidos pelos mesmos e, finalmente, inserindo todas informações referentes ao contato na tabelas *atom_pairs_delaunay* ou *atom_pairs_cutoff*, dependendo do paradigma utilizado.

2.4.1 Filtragem de Arquivos PDB

Arquivos PDB são uma representação de dados de macromoléculas derivadas de processos de resolução de estrutura de proteína usando difração de raios-X, ressonância magnética nuclear, entre outros métodos. Para interpretar esse tipo de arquivo computacionalmente

[5]http://www.wwpdb.org. Último acesso em 17/06/2015
[6]http://www.wwpdb.org/documentation/policy

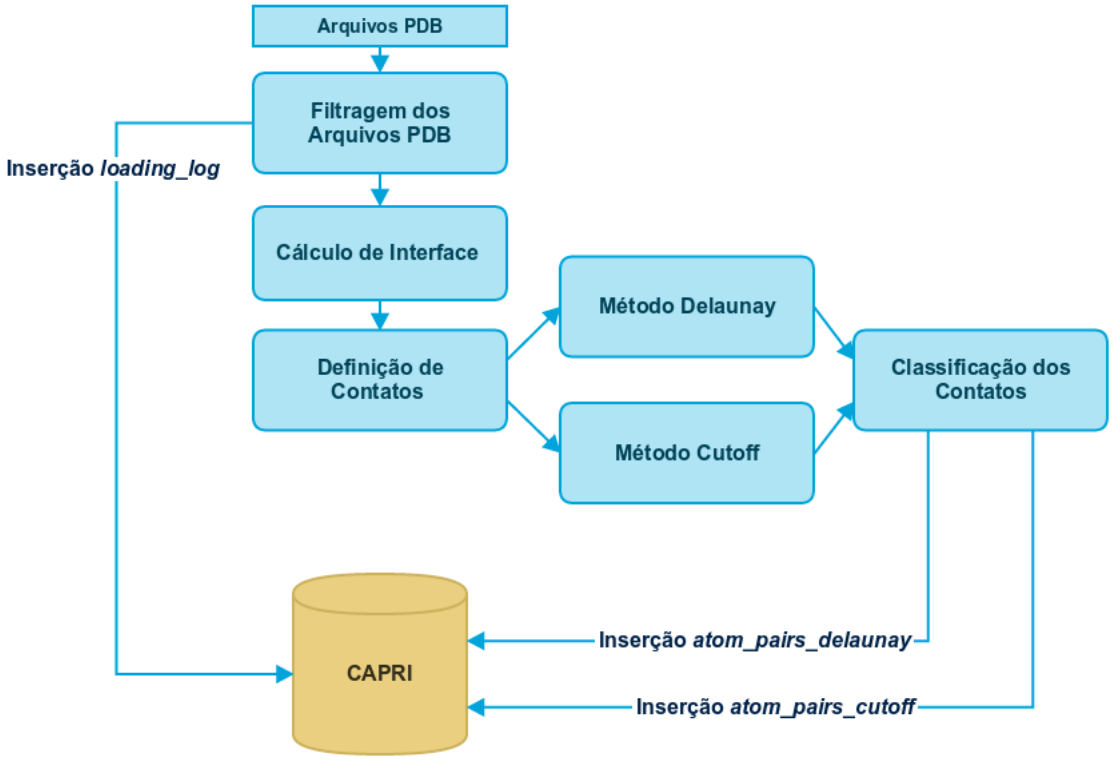


Figura 2.2: Diagramas de fluxo do processo de carga. Em qualquer etapa apresentada, em caso de erro, o mesmo é reportado e inserido na tabela *loading_log*.

utilizamos a biblioteca BioPython [Hamelryck & Manderick, 2003] que permite a criação de um objeto que possua os dados da estrutura da proteína, contendo suas respectivas cadeias, resíduos de aminoácidos e seus átomos com suas posições espaciais. Além disso, foi observado que muitos arquivos podem conter alguns problemas em suas estruturas e é necessário padronizá-los antes do processamento nas próximas fases.

Seguem os tratamentos realizados no arquivo:

- Considerar o átomo de maior ocupância, no caso de haver mais de uma;
- Utilizar apenas o primeiro modelo em caso de arquivos que tenham sido resolvidos por ressonância magnética nuclear;
- Remoção de moléculas de água;
- Remoção de átomos de hidrogênio.

Nesta etapa, aproveitamos para armazenar algumas informações relevantes sobre cada arquivo, uma vez que foram analisados internamente. Essas informações são: quantidade

de cadeias, resíduos e átomos, guardadas dentro da tabela de controle no banco de dados (*loading_log*).

2.4.2 Cálculo da interface proteína-proteína

Uma vez que estamos interessados em avaliar os contatos entre átomos de cadeias polipeptídicas diferentes, ou seja, os contatos inter-cadeia localizados na interface das proteínas, precisamos definir quais são os resíduos que compõem esta região. Esta etapa foi essencial para evitar que resíduos que não estejam na interface sejam avaliados no cômputo dos contatos e reduziu consideravelmente o tempo de computação necessário para este estudo.

Para definir a interface de cada cadeia do complexo proteico utilizamos o método de Lee e Richard Lee & Richards [1971] de acessibilidade ao solvente que consiste em calcular a área de superfície acessível (ASA) de uma proteína em $Å^2$ (angstroms ao quadrado). Para tal usamos o programa NACCESS [Hubbard & Thornton, 1993] que recebe como parâmetro um arquivo PDB e retorna a ASA de cada átomo encontrado.

A seguir, calculamos a ASA do complexo de cada cadeia separadamente e, por fim, identificamos o conjunto de átomos que ganhou acessibilidade ao solvente quando isolada. Se pelo menos um átomo de um determinado resíduo ganhou acessibilidade, consideramos que todos os átomos do respectivo resíduo pertencem à interface. A Figura 2.3 ilustra o resultado desse etapa considerando como exemplo um complexo de duas cadeias polipeptídicas.

Os resíduos que compõem a interface são então marcados para serem analisados na próxima etapa que consiste em determinar os pares de átomos de cadeias distintas que realizam um contato e classificá-los quanto ao tipo de interação estabelecida.

2.4.3 Computação dos contatos

Após a marcação dos resíduos da interface que serão examinados, modelamos a estrutura de cada proteína como um grafo, considerando os átomos como vértices e os contatos entre eles como arestas, e para esta etapa procedemos então utilizando de métodos de dois diferentes paradigmas: Triangulação de Delaunay (*delaunay*) e dependente de distância (*cutoff*). Obviamente, ambos são geométricos e trabalham no espaço tridimensional. Neste estudo, definimos um raio máximo de 7 Å, ou seja, todos os contatos (definidos pelas arestas encontradas) que estiverem a até 7 Å de distância serão computados e armazenados para posterior avaliação. A escolha desta distância remete ao trabalho de Silveira et al. [2009], que observou que a primeira camada de átomos vizinhos se encontra bem delimitada espacialmente com relação às demais camadas. Além disso, essa distância é superior à máxima distância utilizada pela base de dados PICCOLO, que é de 6,05 Å.

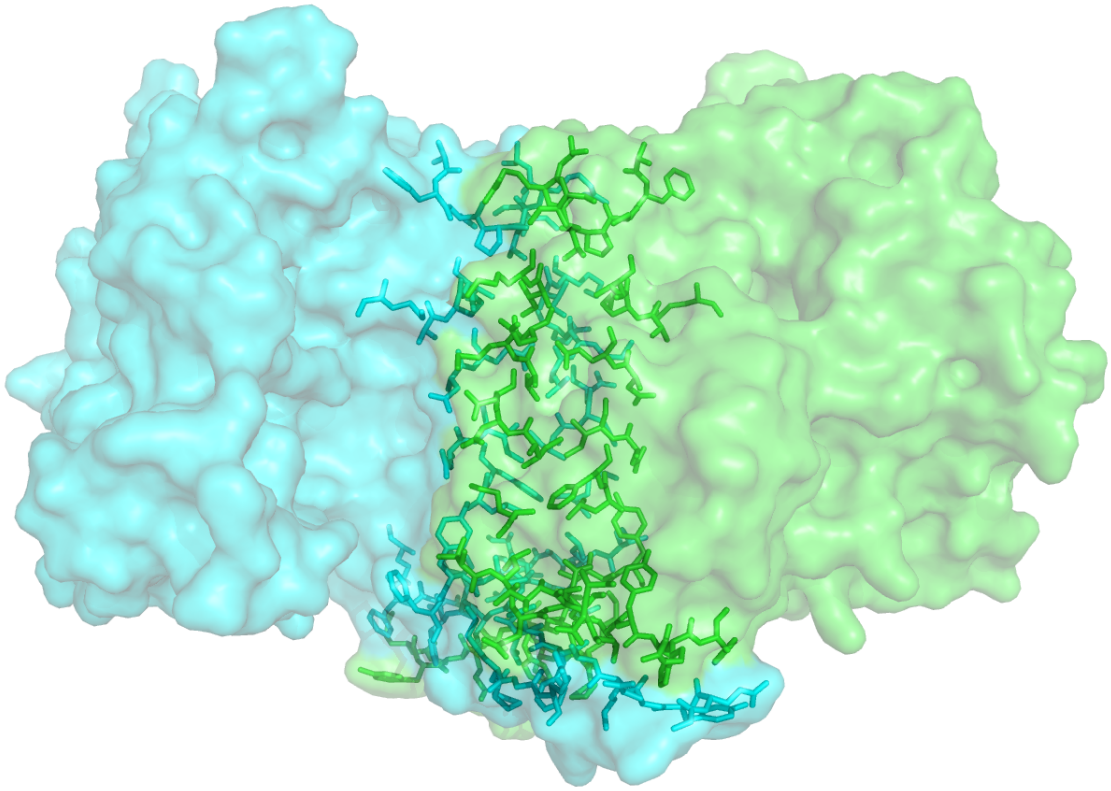


Figura 2.3: Cálculo da interface proteína-proteína. Exemplo de resíduos da interface da estrutura de PDB 1CM7. Os bastões (*sticks*) em verde são resíduos de aminoácidos que compõem a interface da cadeia A e os azuis representam os resíduos da cadeia B

Para o computar os contatos pelo método de *delaunay*, usamos a classe *scipy.spatial.Delaunay* da biblioteca *SciPy*[Jones et al., 2001]. O procedimento dessa classe recebe um conjunto de coordenadas espaciais (*x,y,z*) dos átomos da estrutura e retorna um conjunto de arestas que são representadas como contatos entre dois átomos. Cada aresta foi então analisada, com o objetivo de identificar se os átomos de seus vértices consistem em resíduos válidos, ou seja, um dos 20 resíduos mais comumente encontrados nos seres vivos. Caso contrário, a aresta é desconsiderada do resultado e consequentemente o contato. Além disso, neste primeiro estudo, descartamos também arestas que não possuam os resíduos que façam parte da interface e contatos intra-cadeia, conforme pontuado anteriormente.

O método baseado em distância (*cutoff*) é similar ao *delaunay*, porém não considera qualquer oclusão que venha a ocorrer na varredura. Também foi implementado usando uma classe da biblioteca BioPython[Berman et al., 2000], denominada *NeighborSearch*. O seu uso foi de grande importância para redução do tempo de processamento desta etapa, pois usa um algoritmo de árvore k-D, que cria dados de particionamento de espaço hierárquico de forma eficiente e apesar de ser desenvolvida para linguagem Python, esta classe específica foi implementada na linguagem C++. A classe recebe dois parâmetros: um conjunto de

átomos (sendo esses os átomos dos resíduos das interfaces) e um raio de corte. Para cada átomo passado, a classe varre toda redondeza espacial tendo como origem o centro do átomo, procurando átomos que estejam dentro do raio de corte definido. Assim, consideramos um contato com o átomo referência qualquer outro que esteja dentro do alcance do raio de corte.

A Figura 2.4 mostra um exemplo no qual podemos ver grande variação no número de arestas encontradas pelo paradigma de *cutoff* em contraste com *delaunay*.

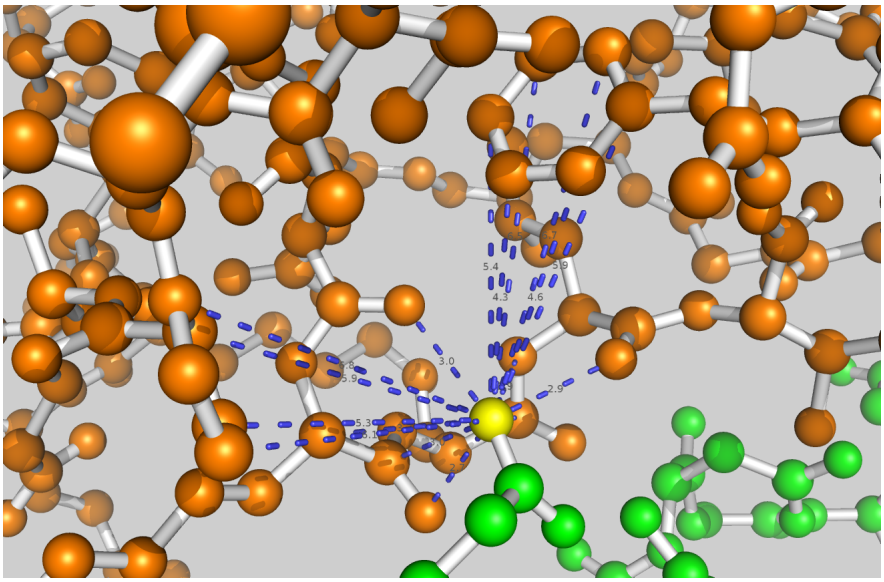
2.4.4 Classificação dos tipos dos contatos

A última fase consiste em determinar qual tipo de contato é estabelecido por um par de átomos. Para que isso fosse possível, cada átomo dos 20 resíduos padrões foi categorizado, quanto às suas propriedades, indicando as possíveis interações que os mesmos podem estabelecer. A Tabela 2.4 apresenta o sumário dessas propriedades:

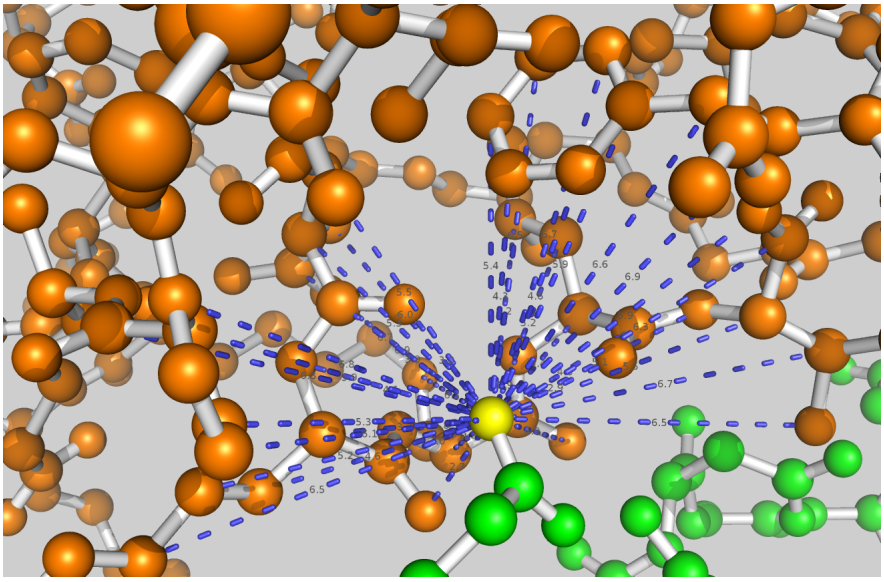
Propriedade	Átomos conforme resíduo
Aceptores	ALA(O), ARG(O), ASN(O, OD1), ASP(O, OD1, OD2), CYS(O, SG), GLN(O, OE1), GLU(O, OE1, OE2), GLY(O), HIS(ND1, NE2, O), ILE(O), LEU(O), LYS(O), MET(O, SD), PHE(O),PRO(O), SER(O, OG), THR(O, OG1), TRP(O), TYR(O), VAL(O)
Aromáticos	HIS(CD2, CE1, CG, ND1, NE2), PHE(CD1, CD2, CE1, CE2, CG, CZ), TRP(CD1, CD2, CE2, CE3, CG, CH2, CZ2, CZ3, NE1), TYR(CD1, CD2, CE1, CE2, CG, CZ)
Doadores	ALA(N), ARG(N, NE, NH1, NH2), ASN(N, ND2), ASP(N), CYS(N, SG), GLN(N, NE2), GLU(N), GLY(N), HIS(N, ND1, NE2), ILE(N), LEU(N), LYS(N, NZ), MET(N), PHE(N), SER(N, OG), THR(N, OG1), TRP(N, NE1), TYR(N, OH), VAL(N)
Hidrofóbicos	ALA(CB), ARG(CB, CG), ASN(CB), ASP(CB), CYS(CB), GLN(CB, CG), GLU(CB, CG), HIS(CB), ILE(CB, CD1, CG1, CG2), LEU(CB, CD1, CD2, CG), LYS(CB, CD, CG), MET(CB, CE, CG, SD), PHE(CB, CD1, CD2, CE1, CE2, CG, CZ), PRO(CB, CG), THR(CG2), TRP(CB, CD2, CE3, CG, CH2, CZ2, CZ3), TYR(CB, CD1, CD2, CE1, CE2, CG), VAL(CB, CG1, CG2)
Ânions	ASP(CG, OD1, OD2), GLU(CD, OE1, OE2)
Cátions	ARG(CZ, NE, NH1, NH2), HIS(CD2, CE1, CG, ND1, NE2), LYS(NZ)

Tabela 2.4: Classificação dos átomos dos 20 resíduos mais comumente encontrados nos seres vivos.

Esta classificação de propriedades atômicas foram baseadas nas mesmas utilizados pela base de dados PICCOLO [Berman et al., 2000]. No escopo inicial do trabalho, utilizamos classificações de outros trabalhos do nosso grupo de pesquisa [Sobolev et al., 1999; de Melo et al., 2007; Gonçalves-Almeida et al., 2012] denomidade SOBOLEV. A Tabela A.1



(a) Arestas utilizando *delaunay*



(b) Arestas utilizando *cutoff*

Figura 2.4: Exemplo da quantidade de arestas encontradas entre os paradigmas *cuttoff* (a) e *delaunay* (b). A imagem é referente à interface do arquivo PDB 1BR8, apresentando um átomo referência (amarelo), que pertence a cadeia I (verde) realizando contatos (azul) com átomos da cadeia L (laranja).

no Apêndice A apresenta uma comparação entre as duas definições. Contudo, obviamente, átomos com classificações distintas entre os métodos resultaram em análises comparativas inconclusivas e adotamos a mesma definição quanto às propriedades.

Quanto às interações possíveis entre pares de átomos, seguem as regras aplicadas:

- Ligação de hidrogênio: aceptor-doador ou doador-aceptor
- Interação hidrofóbica: hidrofóbico-hidrofóbico
- Ponte salina: cátion-ânion ou ânion-cátion
- Empilhamento aromático: aromático-aromático

Vale pontuar que não analisamos interações iônicas repulsivas (cátion-cátion ou ânion-ânion) pois o PICCOLO só contém atrativas.

Após a classificação de cada contato, o mesmo é inserido no banco de dados CAPRI, na tabela referente ao seu método utilizado. Mesmo que um par de átomos não tenham definido um contato pelas suas propriedades, os mesmos ainda sim armazenados, porém com as etiquetas que de contatos marcadas como 0.

2.5 Tabelas derivadas

As tabelas contendo os pares de átomos por contato são imensas com cerca de 367 milhões de registros para o paradigma *cutoff*, 209 milhões para o PICCOLO e 76 milhões para o *delaunay*. Lidar com essa base por si só já é um desafio computacional. Para melhorar o desempenho das consulta nessas bases, criamos índices no campos que são comumente utilizados para consulta. Além disso, o mecanismo de armazenamento(*"Storage Engine"*) utilizado nas tabelas foi o MyISAM. Este tipo de mecanismo é recomendado quando se fazem mais consultas do que modificações nas tabelas, exatamente o caso do CAPRI que é um banco de dados analítico e estático.

Para a etapa de análise dos dados e para melhorar o desempenho das ferramentas de visualização de dados, foi necessário projetar e implementar tabelas com dados extraídos das três tabelas principais (*atom_pairs_piccolo*, *atom_pairs_delaunay*, *atom_pairs_cutoff*). Estas novas tabelas possuem dados agregados por distância, armazenando tanto soma de quantidade de contatos como mediana. São elas:

- median_pdb_contact
- sum_pdb_contact

- sum_res_atom_contact

A tabela *median_pdb_contact* contém dados quanto a mediana de cada paradigma e tipo de contato agrupadas por cada PDB e intervalo de distância. O propósito para a construção desta tabela foi para possibilitar a implementação do teste de hipótese de Wilcoxon que será abordado na Seção 2.7. As colunas desta tabela possui os valores das medianas de cada paradigma (ao todo 3) por cada interação abordada (ao todo 4), totalizando 12 colunas, agrupadas por PDB e distância. Para melhor explicarmos, tomemos os seguintes dados como exemplo: o PDB 1A0D possui 9 ligações de hidrogênio quando se usa o paradigma *piccolo* à distância de 3,3 Å. Quando verificamos na tabela *atom_pairs_piccolo*, que contém as interações do paradigma *piccolo*, verifica-se a somatório de ligações de hidrogênio para cada intervalo de distância dentro de 3,3 Å, ou seja, distâncias entre 3,25 a 3,34 quando aumentamos uma casa decimal, temos os seguintes resultados de intervalo de distância e somatória de contatos de ligações de hidrogênio respectivamente: 3,25 = 2; 3,26 = 1; 3,27 = 2; 3,28 = 1; 3,32 = 2; 3,33 = 1, totalizando 9 interações. Então temos que a mediana de ligações de hidrogênio no paradigma *piccolo* para o PDB 1A0D no intervalo de distância de 3,3 Å é igual a 1,5.

A tabela *sum_pdb_contact* possui a mesma estrutura da mediana, contendo também 2 colunas agregadas, sendo elas os código do arquivo PDB e um intervalo de distância (com 1 casa decimal) e 12 colunas quantitativas que armazenam desta vez a somatória de contatos registrados por paradigma e por tipo de contato.

Por último temos a tabela de soma por resíduo e átomo (*sum_res_atom_contact*). Nela os dados não estão agregados por PDB (como nas tabelas anteriores) e sim por resíduo, átomo e intervalo de distância, nesta ordem, utilizando desta vez 3 colunas de agregação. Este agrupamento permite uma visão mais global da base de dados permitindo uma análise mais abrangente do comportamento dos contatos em cada intervalo de distância. Esse agrupamento pode ainda ser usado em visualizações agregadas por nível de resíduo ou ate mesmo a nível atômico. Na ferramenta desenvolvida, é possível analisar os dados nessas diferentes granularidades. Contudo não percebemos variações significativas por nenhum resíduo ou átomo particular. Contudo, não nos aprofundamos nessas análises por questão de tempo, mas como trabalho futuro, pretendemos nos aprofundar nessas análises no nível de resíduo e átomo.

2.6 Ferramenta de visualização para análise comparativa dos dados

Desenvolvemos uma página web para apresentar visualização dos dados de forma interativa. Com isso facilitamos a exploração da base e a realização de análises comparativas, objeto de estudo desse trabalho. Ela foi implementada usando D3.js, que é um biblioteca de *JavaScript* que permite criar gráficos interativos e dinâmicos, manipulando imagens do tipo *SVG* (*Scalable Vector Graphics*). Este tipo imagem permite criar desenhos e gráficos utilizando formas vetoriais, seja de forma estática, dinâmica ou animada, e sua grande vantagem é que ela não perde qualidade ao ser ampliada. A única desvantagem observada quando utiliza-se a biblioteca D3.js é o custo de aprendizado, pois é necessário conhecer alguns conceitos intermediários de *JavaScript*.

Utilizando o D3.js, criamos três gráficos de linha, contendo informação sobre as distribuições das frequências dos contatos a cada distância. Cada linha representa um dos paradigmas abordados, com exceção do terceiro gráfico que aborda um par de paradigmas e ilustra os P-valores resultantes do teste de hipótese utilizado. A Figura 2.5 apresenta a interface da página criada, contendo os gráficos, como caixas de combinação (*combo box*) e de seleção (*checkbox*) para variação dos parâmetros.

É importante ressaltar que o diagrama de Venn apresentado junto ao terceiro gráfico de linha é meramente ilustrativo e não é utilizado para representar alguma grandeza especifica. O diagrama representa somente a legendas das cores dos paradigmas e seus pares (interseções nas cores ciano, amarelo e magenta), apresentados no terceiro gráfico de linha.

A página exibe as análises de um único tipo de contato por vez e é possível alterar a visualização por contato na caixa de combinação de respectivo nome. Além disso, a página possui opções de visualização de contatos agregando as informações por resíduo ou por átomos, porém o filtro se aplica somente ao primeiro gráfico, uma vez que os dados dos outros dois estão agregados por PDB. É possível ainda realizar um *zoom* sobre os gráficos, utilizando os campos do *Intervalo de distância* onde o usuário define a distância inicial e final. Essa funcionalidade se mostrou muito útil quando é preciso analisar uma região local e além disso ela permite ver dados que não são possíveis quando analisamos o intervalo completo de distância de 0,0 a 7,0 Å, como mostra a Figura 2.6. É possível perceber que ao visualizarmos o gráfico à nível global (retângulo tracejado menor) entre 2 a 2,5 de distância (Å), as curvas azul e vermelha parecem idênticas, porém, quando se aproxima (retângulo tracejado maior), é possível verificar que elas não se tocam em nenhum momento. Por fim, temos uma caixa de combinação apenas para suavizar as curvas, tornando a visão um pouco mais agradável por reduzir possíveis "serrilhados", porém um pouco menos realista e precisa

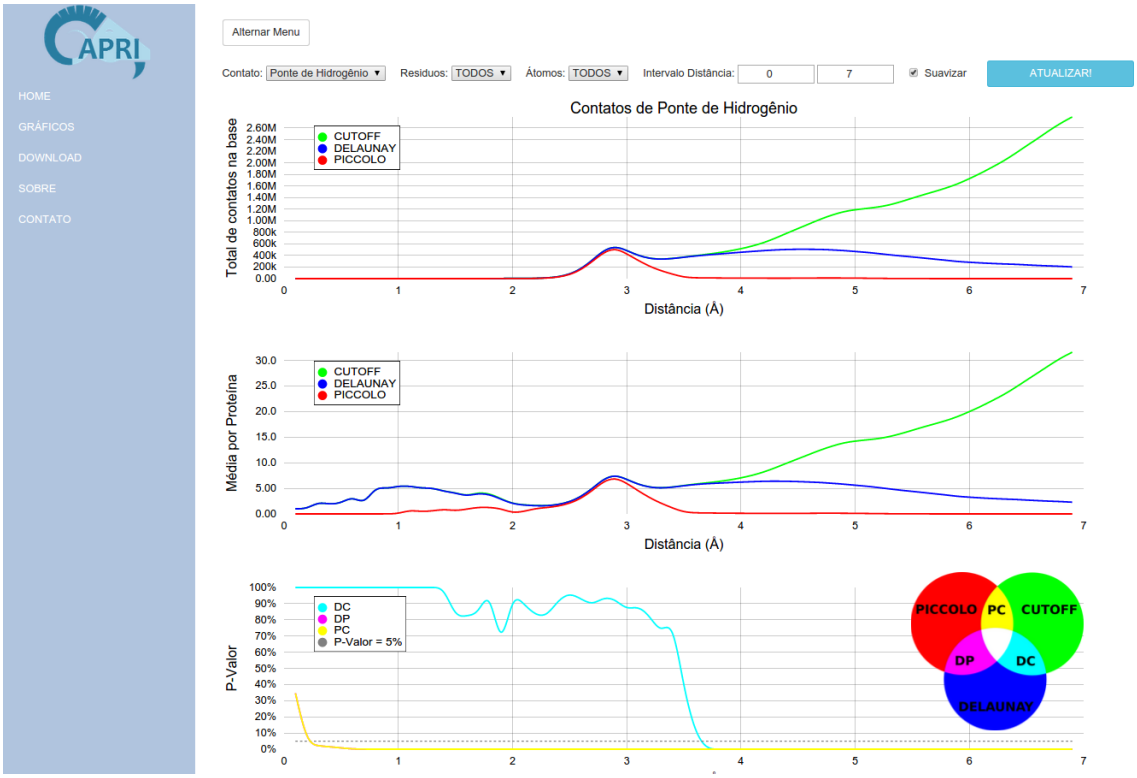


Figura 2.5: Interface da página web criada para visualização dos resultados.

dos dados originais.

Por ser interativo, qualquer mudança nos parâmetros não atualiza a página como um todo, ocorrendo apenas uma animação de interpolação das linhas, ajustando o valores para os argumentos selecionados.

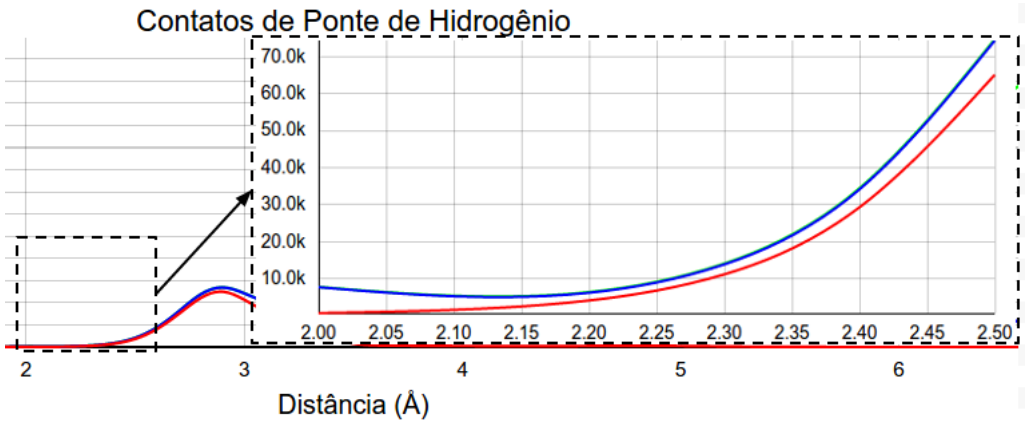


Figura 2.6: Exemplo de uso da funcionalidade de *zoom* para observações mais detalhadas de um determinada região no gráfico.

Os dados dos gráficos estão associados à arquivos do tipo TSV (*Tab-Separated Values*). Estes arquivos textos foram criados a partir de *scripts* em Python que realizam consultas nas

tabelas e organizam o formato do texto da maneira que D3.js possa lê-los e consequentemente, construir o gráfico. Esta ação não é feita em tempo real ou por demanda do usuário, pois a página não usa um servidor para criar uma conexão entre o banco de dados sendo estes arquivos estáticos e pré-processados.

2.7 Metodologia para análise estatística dos dados comparativos

Observando os gráficos, podemos tentar identificar visualmente o ponto a partir do qual uma determinada curva de um paradigma começa a divergir das outras. Entretanto, como explicado na Seção 2.6 sobre a justificativa para criarmos a funcionalidade, conforme figura 2.6, dependendo da resolução em que estejamos analisando os gráficos, pode ser impossível visualizar as diferenças existentes. Obviamente que não é desejável que a estimativa das semelhanças e diferenças entre essas curvas seja feita de forma pouco rigorosa visto que essa divergência é o principal ponto a ser investigado nesse trabalho. Assim, buscamos uma metodologia formal e sistemática para identificar o ponto onde as curvas divergem. Como não foi observado um distribuição Normal nos resultados, não caberia a escolha de um teste de hipótese paramétrico e com isso decidimos utilizar o teste de postos com sinais de Wilcoxon [Wilcoxon, 1945] que é um método não-paramétrico para comparação de duas amostras pareadas com o objetivo de verificar se elas são iguais ou não. O seu funcionamento consiste em calcular as diferenças entre cada elemento pareado das amostras, sendo possível três condições de resultados destas diferenças: positivo (+), negativo (-) ou zero (0). Após este cálculo, as diferenças entre os resultados para cada par de dados são ordenadas pelo seu valor absoluto (desconsiderando o sinal), substituindo-se então os valores originais pelo posto ou *ranking* que ocupam na escala ordenada. O teste de hipótese para avaliar a igualdade entre as amostras é então baseado em uma estatística de teste levando em conta a soma dos postos multiplicados pelo sinal da diferença que contém valores positivos e negativos. Convencionalmente, se usa um nível de significância de 5% (α = 0,05), um P-valor maior que α_1 indica que as amostras são iguais.

Para aplicar o teste de Wilcoxon aos pares de paradigmas usamos a tabela *median_pdb_contat*. Como já explicado, ela possui a mediana dos paradigmas e tipos de contatos agregados por PDB e distância (em Å). Sendo assim, para cada par de paradigmas aplicados ao teste são filtrados para cada intervalo de distância (variando de 0,0 até 7,0 Å) e tipo de contato (ligação de hidrogênio, interação hidrofóbica, ponte salina e empilhamento aromático). Os resultados do teste são apresentados no terceiro gráfico das análises, definindo um sigla para cada par, sendo: DC = *delaunay-cutoff*, DP = *delaunay-piccolo* e PC =

piccolo-delaunay.

2.8 Dificuldades encontradas

Muitas dificuldades foram encontras na condução deste estudo, porém achamos importante destacar uma em particular para que outros pesquisadores não passem pelo mesmo problema. A tabela original da base de dados PICCOLO foi desenvolvida com o campo de distância tendo como tipo *ponto flutuante* (*Float*). Este tipo de campo é comumente usado no campo científico pois podem armazenar muitas casas decimais, permitindo maior precisão em cálculos complexos. Para este trabalho, o campo de distância não precisa de tanta precisão. Acreditamos que entre 2 e 3 casas decimais seriam mais que suficientes. Como usamos arredondamento de distância para reduzir as casas decimais de algumas das tabelas, ao analisarmos os valores prévios nos gráficos observou-se padrões ou oscilações periódicas sistemáticas nas linhas do gráfico. Inicialmente achamos esses padrões sem nenhuma explicação aparente. Foi necessário muito tempo e esforço até descobrirmos que o problema não estava nos dados, mas no campo, ao aplicar o arredondamento. O que ocorria era que a redução de casas decimais não era feita de forma justa. Quando se arredonda um valor, as distribuições de intervalo deviam ser iguais, e para campos do tipo *ponto flutuante* isso não ocorreu devido ao fato deste não definir casas decimais fixas, sendo a sua quantidade determinada pela capacidade do processador do computador. Tendo como exemplo o valor 2,5, temos um intervalo de 2,45 a 2,54 (10 valores). Quando se usava este mesmo valor, sendo ele armazenado como *ponto flutuante*, tínhamos 2,45 a 2,55 (11 valores) ou até mesmo em alguns casos 2,46 a 2,54 (9 valores), gerando assim resultados de curvas que oscilavam em cada intervalo de distância. Para resolver este problema, convertemos todos os campos de todas as tabelas que usam distância para o tipo *decimal*, resolvendo assim o problema, uma vez que nesse cenário as casas decimais são fixadas e não "flutuam"como o próprio nome do campo de *ponto flutuante* indica.

Capítulo 3

Resultados e Discussões

A principal contribuição deste trabalho foram o desenvolvimento de uma base de dados com contatos em interfaces proteína-proteína com o uso de diferentes paradigmas bem como a análise comparativa de contatos obtidos pelos diferentes métodos a diferentes distâncias para ligações de hidrogênio, interações hidrofóbicas, pontes salinas e empilhamentos aromáticos. Essa pesquisa gerou alguns artefatos que serão disponibilizados publicamente na web. Dentre eles estão a base de dados CAPRI e o código-fonte da aplicação utilizada para carregar o banco de dados, descritos no Capítulo 2. Iniciamos esse capítulo apresentando os artefatos gerados e finalizamos com as discussões referentes às análises comparativas realizadas.

3.1 Artefatos produzidos

Como artefatos produzidos disponibilizamos o banco de dados e o código-fonte em Python usado para carregar a base de dados. Estes artefatos estão disponíveis em *http://homepages.dcc.ufmg.br/~pmartins/capri1/download*.

3.1.1 Base de dados

A base de dados CAPRI possui o tamanho de 53 GigaBytes (7,3 GigaBytes no formato compactado). Ela foi armazenada como um arquivo de *script* (*.sql*) e para carregá-la após o *download* basta realizar o procedimento de restauração no banco de dados desejado. Segue exemplo de comando para restauração utilizando SGBD MySQL:

mysql -u <*usuario*> -p*senha* capri < capri1.0.sql

Lembrando que é necessário associar o *script* a uma base de dados já existente.

3.1.2 Código-fonte

O código-fonte que foi usado para carregar o banco de dados implementado em Python possui um arquivo *main.py* que representa o programa principal. Além disso, um pacote (capri) foi criado para armazenar dois arquivos: *util.py* e *mysql.py*, ambos contendo funções utilizadas pela programa. O primeiro contêm os módulos criados para ler, filtrar, definir interfaces, etc, a partir de arquivos PDBs. O segundo contêm funções de leitura e escrita com o banco de dados. Além disso, é necessário configurar adequadamente o arquivo *mysql.ini*, localizado na pasta *loading*, para criar uma comunicação com a base de dados. Por fim, disponibilizamos um *script* (*.sql*) na pasta *sql* chamado *database_capri_min.sql*, que cria a base de dados CAPRI com mínimo de dados necessário para que o usuário possa replicar o que foi realizado neste trabalho.

Por se tratar da versão inicial (1.0), o programa não possui ainda uma usabilidade amigável, necessitando de estruturas de diretórios fixas para que funcione com exatidão. O principal interesse em disponibilizar o código-fonte parte da intenção de deixar público os procedimentos realizados para adquirir as informações do banco de dados o mesmo é livre para uso, modificação ou replicação. Temos a intenção de melhorar o código para que seja mais amigável e incrementar com novas funcionalidades que estaremos disponibilizando ao longo do tempo em novas versões.

3.2 Análise dos resultados

Discutimos a seguir os resultados obtidos na análise comparativa dos quantitativos de contatos obtidos quando consideramos diferentes distâncias para o cálculo de contatos em interfaces protéina-protéina.

Fizemos basicamente três tipos de análises, cada qual ilustrada em um gráfico de linhas. O primeiro gráfico consiste na distribuição do número total de contatos obtidos na base de dados inteira quando se considera distâncias no intervalo de 0 a 7 Å com os três diferentes paradigmas estudados, a saber: *cutoff*, *delaunay* e *piccolo*. A segunda análise é análoga à primeira mas retrata a distribuição do número médio de interações por complexo proteína. Os eixos X de todos os gráficos representam a distância de cada par de átomos que estabelecem um contato e os eixos Y variam conforme o gráfico, sendo: o primeiro representado pela quantidade total de contatos obtidos na base; o segundo a média por proteína. O que observamos com relação à essas duas análises é que os perfis das curvas total e média são comparáveis exceto no início (a curtas distâncias) onde se observam oscilações na curva média que se devem a valores extremos que influenciam na média.

Inspirados no trabalho anterior do nosso grupo de pesquisa Silveira et al. [2009], o

ponto focal das nossas análises é o ponto onde as curvas referentes aos três paradigmas se diferenciam, o que indica um possível valor de distância mais adequado para determinado tipo de contato. No trabalho de Silveira, os autores concluem que o ponto de 7,5 Å representa o cutoff a partir do qual o método de distânca começa a apontar contatos falsos positivos devido principalmente à contatos oclusos por outros átomos.

Uma contribuição deste trabalho e uma das diferenças entre este trabalho e o de Silveira é que propomos a identificação desse ponto de divergência através de um teste estatístico e não apenas visualmente. Esse processo foi explicado em detalhes na seção 2.7. Em linhas gerais, computamos os p-valores para os pares de métodos e plotamos na terceira curva o seu valor para as diferentes distâncias consideradas na análise comparativa. Esses resultados são apresentados na terceira curva (terceiro tipo de análise proposto). Mais uma vez, o eixo X representa a distância de cada par de átomos que estabelecem um contato e o Y apresenta o resultado do teste de hipótese de Wilcoxon que avalia o quanto dois métodos são semelhantes à uma determina distância específica. O nível de significância escolhido foi 5% (α = 0,05), que é o valor padrão praticado em testes de hipótese e significa que em qualquer momento da curva onde a mesma esteja abaixo de 5% implica que a hipótese de igualdade entre os paradigmas foi rejeitada, implicando que os métodos são distintos naquele ponto.

Assim, para cada tipo de contato estudado, apresentamos um conjunto de três gráficos que devem ser analisados em conjunto. O principal objetivo desses gráficos é destacar as semelhanças e diferenças entre as curvas das distribuições dos três paradigmas que representam a quantidade de contatos obtidos à diferentes distâncias, em seu total e média. Conforme explicado anteriormente na Seção 2.7, trabalhamos com confiança de 95% e, para tanto, consideramos que um par de métodos começa a se diferenciar quando a respectiva curva de p-valores cai abaixo de 5% e converge para valores abaixo deste limiar.

3.2.1 Ligações de hidrogênio

A Figura 3.1 ilustra os resultados da análise do quantitativo de contatos do tipo ligação de hidrogênio obtidos pelos três diferentes paradigmas estudados. No primeiro gráfico, apontamos como principal padrão visual a ocorrência de um pico começando no marco de distância de aproximadamente 2,2 Å. A partir desse ponto, notamos que a curva referente ao método *piccolo* (em vermelho) começa a se mostrar progressivamente diferenciada em comparação com as demais (*cutoff* e *delaunay*), decaindo gradativamente até por volta de 3,6 Å. Ainda pela análise visual, notamos que a curva que ilustra o método *cutoff* (verde) explode quando a distância cresce enquanto o mesmo não se verifica com a curva do método de *delaunay* (azul).

Uma análise semelhante da diferença entre as curvas de *cutoff* e *delaunay* foi objeto de

estudo de Silveira et al. [2009] e colaboradores mas aquele estudo foi um pouco diferente. Basicamente, eles analisaram o ponto de divergência entre as curvas, mas as curvas do estudo deles retratavam o número de vizinhos (ou contatos) a uma certa distância sem considerar o tipo de interação e usando um centróide que representava o resíduo e não em nível atômico como apresentamos nesse estudo. Assim, os valores obtidos por eles não são comparáveis aos nossos, os estudos são apenas qualitativa e conceitualmente relacionados.

As curvas de *cutoff* (verde) e *delaunay* (azul) se diferenciam como era de se esperar, pois a curva do método baseado em distância reflete o fato que quanto maior a distância mais contatos serão retornados pelo método o que não ocorre com o método de *delaunay* que pegará apenas a primeira camada de vizinhos, não apresentando o inconveniente de retornar uma enorme quantidade de contatos falsos positivos.

O mais interessante na análise das ligações de hidrogênio vem na diferença entre as curvas do *piccolo* e das demais. Observe que a curva do *piccolo* (vermelha) decai após cerca de 3,6 Å e a partir desse valor o número de contatos obtidos é desprezível. Esse seria o melhor resultado considerando a natureza das ligações de hidrogênio e o método obtém essa precisão por considerar a angulação entre os átomos.

Por fim, vale a pena destacar o ponto 3,9 Å que é o ponto de máximo da curva do método *piccolo* e indica o valor mais frequente ou seja o limiar no qual temos o maior número de ligações de hidrogênio calculadas.

3.2.1.1 Exemplo

A Figura 3.2a ilustra um hidrogênio (branco) ligado ao um nitrogênio (azul) doador que está a 3,0 Å de distância de um oxigênio (vermelho) realizando uma ligação de hidrogênio (aresta amarela), se não considerada a angulação. Como o paradigma *delaunay* por se só não possui este critério, o contato é dado como verdadeiro. Já no *piccolo*, a distância de 3,0 Å determina a seguinte regra: o ângulo entre o hidrogênio, o doador e o aceptor deve ser inferior que 90 °, o que não ocorre nesse exemplo no qual o ângulo observado é de 109,5 °.

O segundo exemplo, conforme Figura 3.2b, mostra claramente uma oclusão gerada pelo Cα (verde), impedindo a formação de uma ligação de hidrogênio (tracejado amarelo) entre um nitrogênio (azul) e um oxigênio (vermelho) quando se usa o paradigma *delaunay*, porém este contato é valido utilizando *cutoff* se a distância de 4,5 Å fosse considerada válida.

3.2.2 Interações hidrofóbicas

Com relação às interações hidrofóbicas, as análises que podem ser feitas são análogas às pontes de hidrogênio. Contudo, a análise das curvas e, tendo em vista que interações hidrofóbicas não tem dependência física com o critério de angulação, o que se observa é que

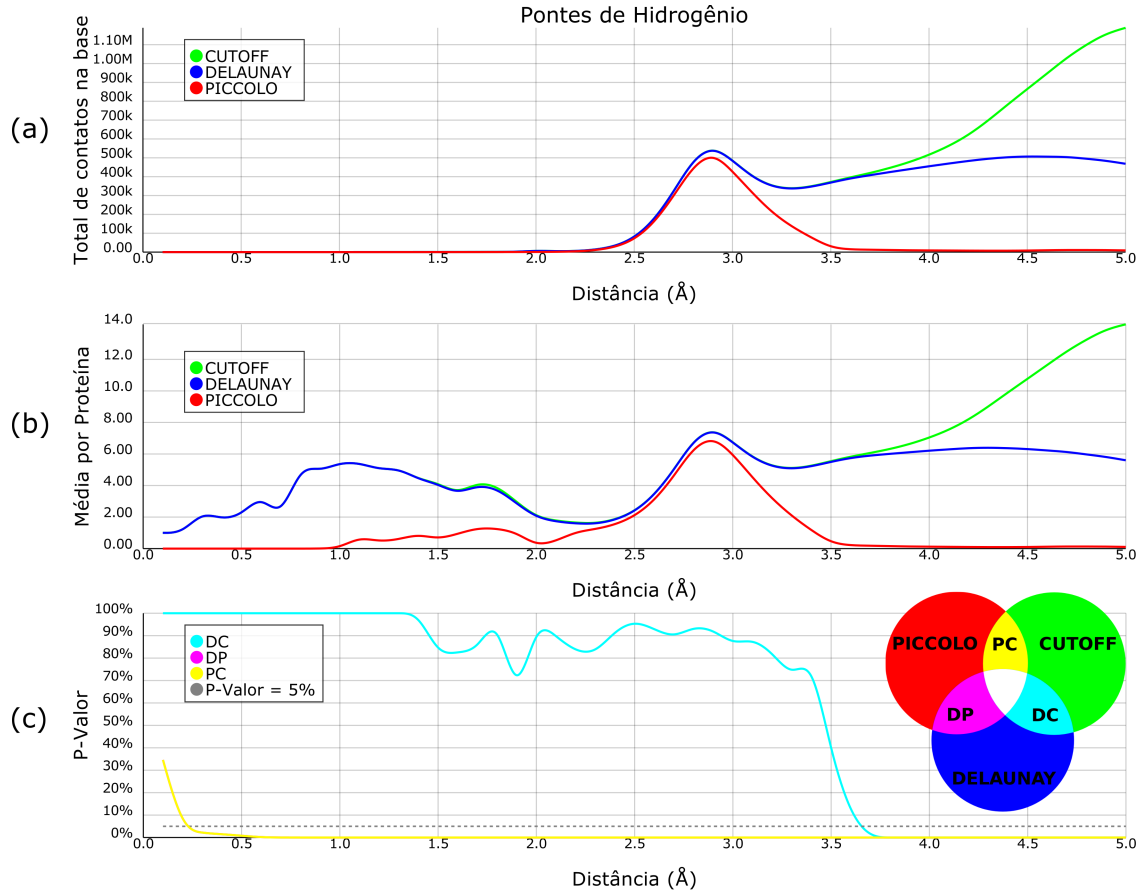
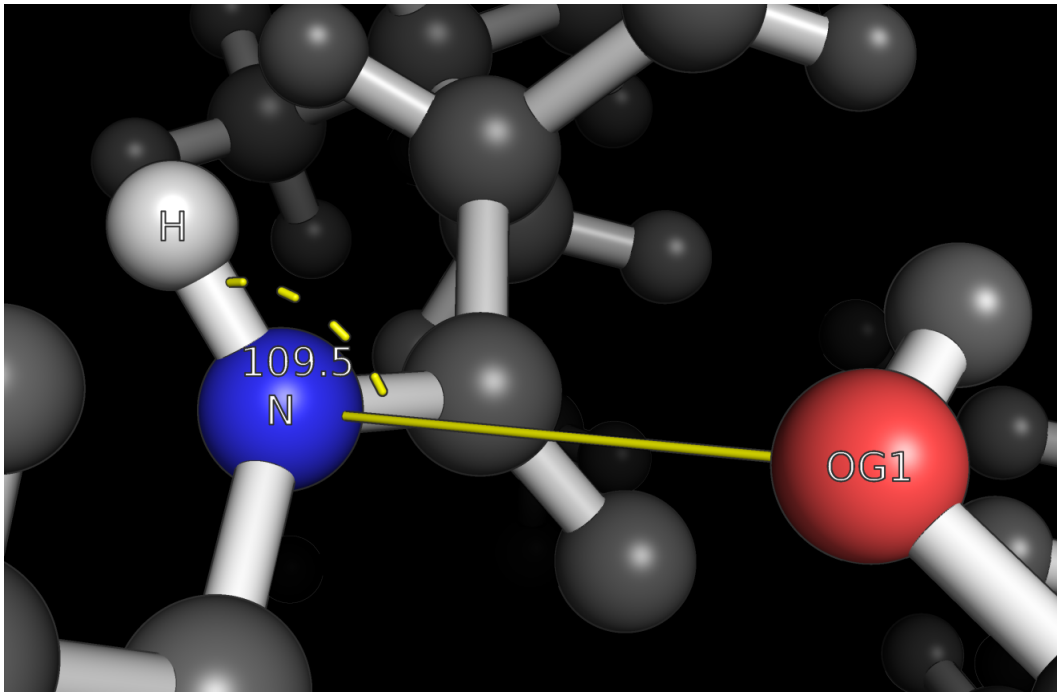


Figura 3.1: Análise comparativa dos paradigmas no cálculo de ligações de hidrogênio.

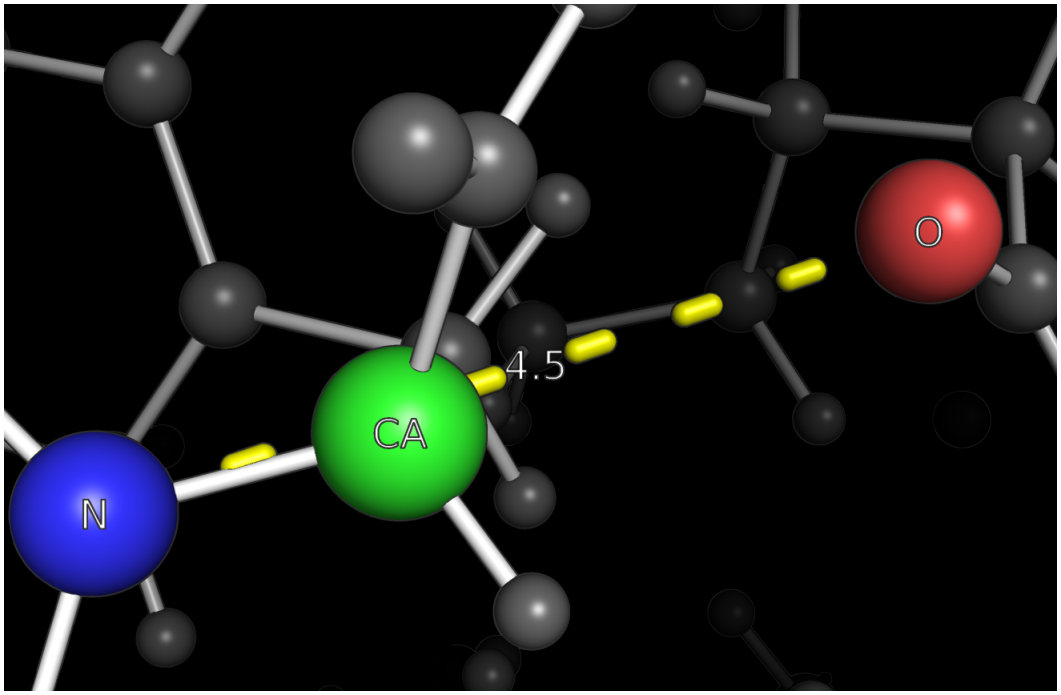
a principal diferença se dá pelo critério de oclusão. Em outras palavras, o método baseado em *cutoff* e o *piccolo* não consideram oclusão e são bastante comparáveis, até porque são calculados de forma semelhante. O principal ponto de foco aqui está em 3,8 Å quando os métodos puramente baseados em distância divergem e começam a trazer um considerável número de falsos positivos. Isso nos leva a crer que, ao menos em interfaces proteína-proteína, a uma distância igual ou menor que esse limiar (3,8 Å) o número de interações hidrofóbicas tidas como falso positivas é insignificante enquanto acima desse valor, é preciso considerar critérios de oclusão para garantir a corretude dos contados prospectados. Um outro ponto interessante aqui é o 4,2 Å que é o ponto de máximo da curva do método *delaunay* e indica o valor mais frequente ou seja o limiar no qual temos o maior número de contatos hidrofóbicos prospectados.

3.2.2.1 Exemplo

No cômputo das interações hidrofóbicas não existe critério de angulação. Apesar disso, verificamos alguns contatos próximos com distância por volta de 4,2 Å entre *delaunay* e os



(a) Comparação de ligação de hidrogênio entre os paradigmas de *piccolo* e *delaunay*. Diferença de critério de angulação localizada no arquivo de PDB 1A1U: TRP29:C e GLU31:A.



(b) Comparação de ligação de hidrogênio entre os paradigmas de *delaunay* e *cutoff*. Diferança por oclusão localizada no arquivo de PDB 1BVG: ILE93:A e o PHE99:B.

Figura 3.2: Exemplo de comparação de ligações de hidrogênio entres os paradigmas estudados.

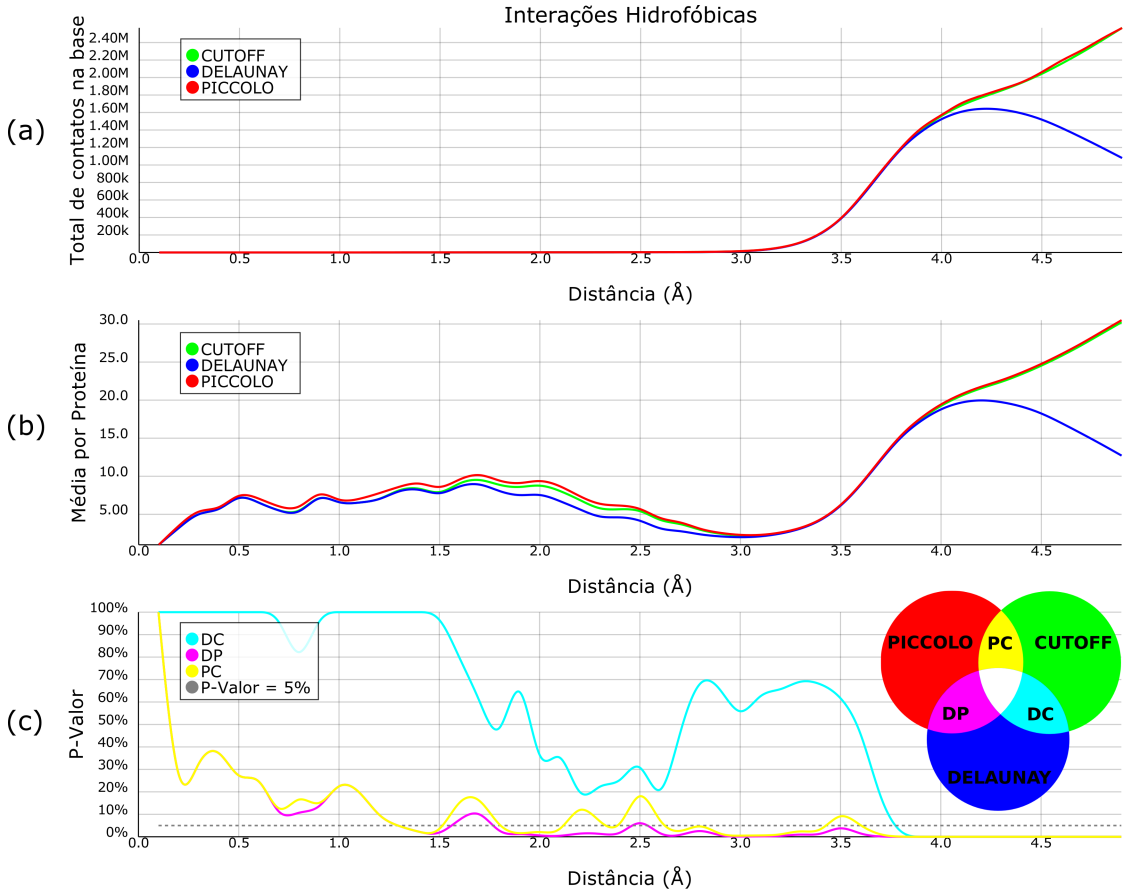
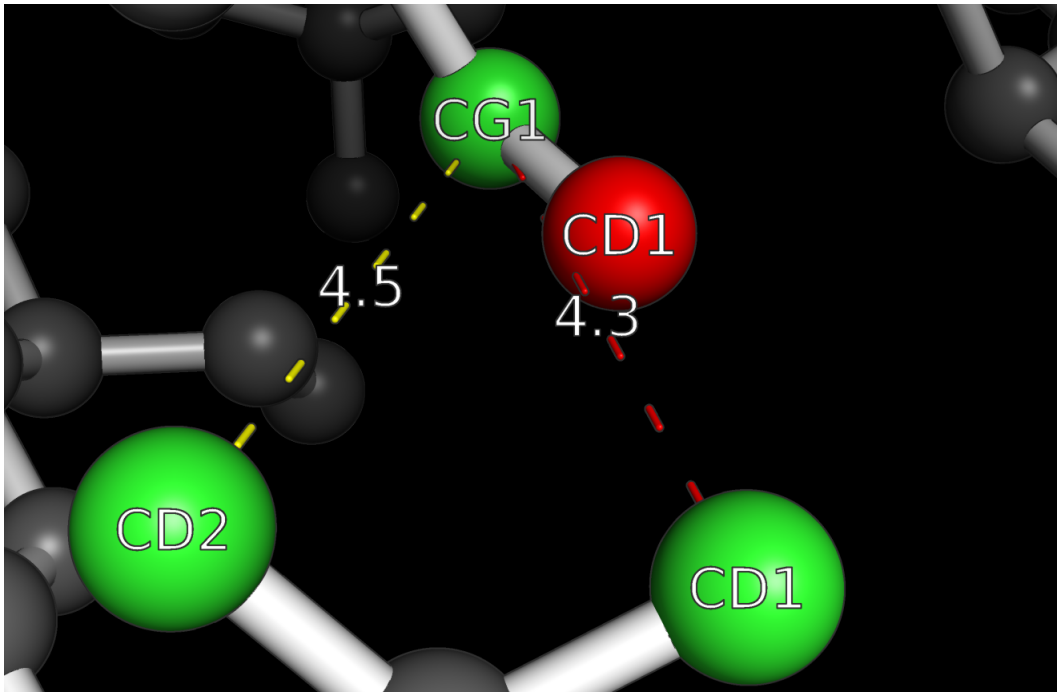


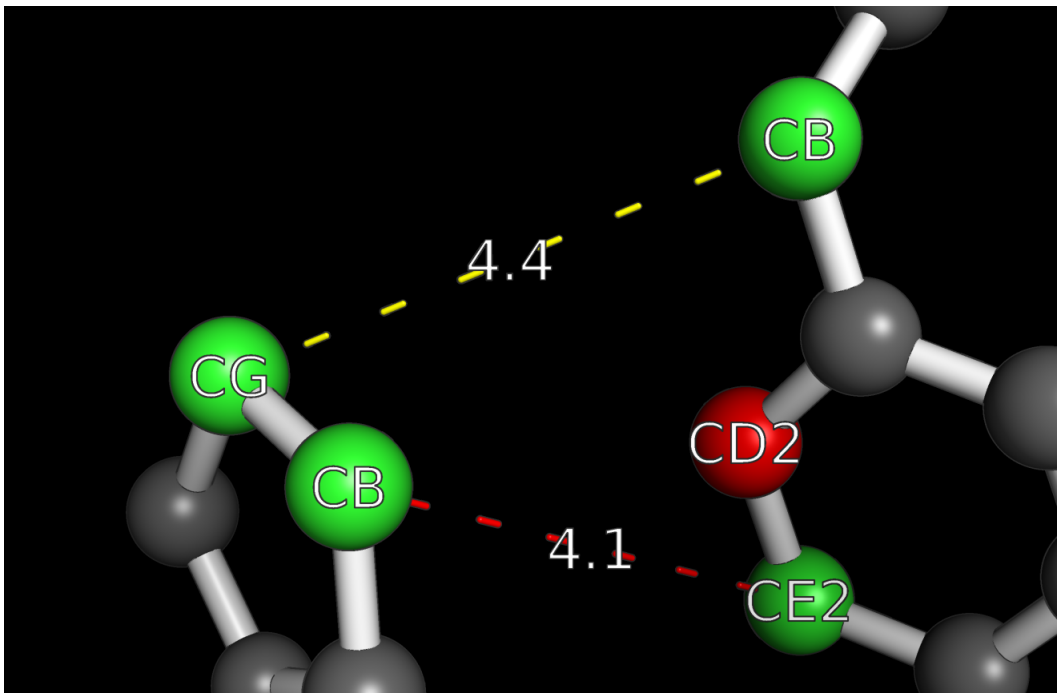
Figura 3.3: Análise comparativa dos paradigmas no cálculo de interações hidrofóbicas.

demais paradigmas para avaliar as diferença nas curvas neste ponto. A Figura 3.4a representa uma comparação entre *piccolo* e *delaunay*. O contato entre o CD2 da LEU com o CG1 da ILE (tracejado amarelo) realiza um interação hidrofóbica tanto no *piccolo* como no *delaunay*. Já no contato entre o CD1 da LEU (verde) com o CG1 da ILE, o CD1 (vermelho) cria um oclusão e somente o paradigma *piccolo* aceita como um interação hidrofóbica (tracejado vermelho).

Na Figura 3.4b temos uma comparação entre *cutoff* e *delaunay*. Percebe-se que não há oclusão entre o contato de CG da PRO com CB da TYR (tracejado amarelo), sendo esta aresta aceita tanto no *delaunay* como no *cutoff*. Ao observarmos o contato entre CB da TYR com CE2 da PRO (tracejado vermelho) não se detecta trivialmente uma oclusão, porém a geometria realizada na Triangulação de Delaunay cria somente uma aresta entre CB da PRO e CD2 (vermelho) da TYR, e com isso o paradigma *delaunay* não considera a interação hidrofóbica entre CB da TYR com CE2 da PRO (tracejado vermelho).



(a) Comparação de interação hidrofóbica entre *piccolo* e *delaunay*. PDB 1AFO: LEU75:A e ILE7:B.



(b) Comparação de interação hidrofóbica entre *cutoff* e *delaunay*. PDB 4GBQ: TYR7:A e PRO2:B.

Figura 3.4: Exemplo de comparação de interações hidrofóbicas entres os paradigmas estudados.

3.2.3 Pontes salinas

De forma similar ao que ocorre com as interações hidrofóbicas, as pontes salinas também não são fisicamente dependentes da angulação de forma que o critério mais importante

para se evitar a recuperação de falsos positivos é mesmo a oclusão implementada através do método de *delaunay*. Dessa forma, nota-se com essa análise grande similaridade entre os métodos paseado em *cutoff* e *piccolo*, como esperado. O método de *delaunay* por sua vez se diferencia significativamente com 95% de significância dos outros a partir de 3,1 Å, ou seja, a partir desse ponto pode-se esperar um crescente número de falsos positivos devido à oclusão também.

Essa curva apresenta um perfil bastante curioso e cujo significado ainda não fomos capazes de elucidar. Note que a curva não é monomodal mas bimodal no caso do método baseado em oclusão e trimodal nos métodos que não a consideram. O primeiro modo tem seu pico em torno de 2,9 Å, o segundo em torno de 3,6 Å e o último nas proximidades de 4,9 Å. O primeiro modo tem seu pico similar para os três métodos mas outro fato curioso é que o segundo pico é um pouco deslocado no método baseado no delauney (3,55 Å) com relação aos outros dois (3,65 Å). Não entendemos por que isso ocorre e nem o significado biológico desses vários modos na curva de distribuição sendo essa uma direção para futura investuigação e aperfeiçoamento desse trabalho.

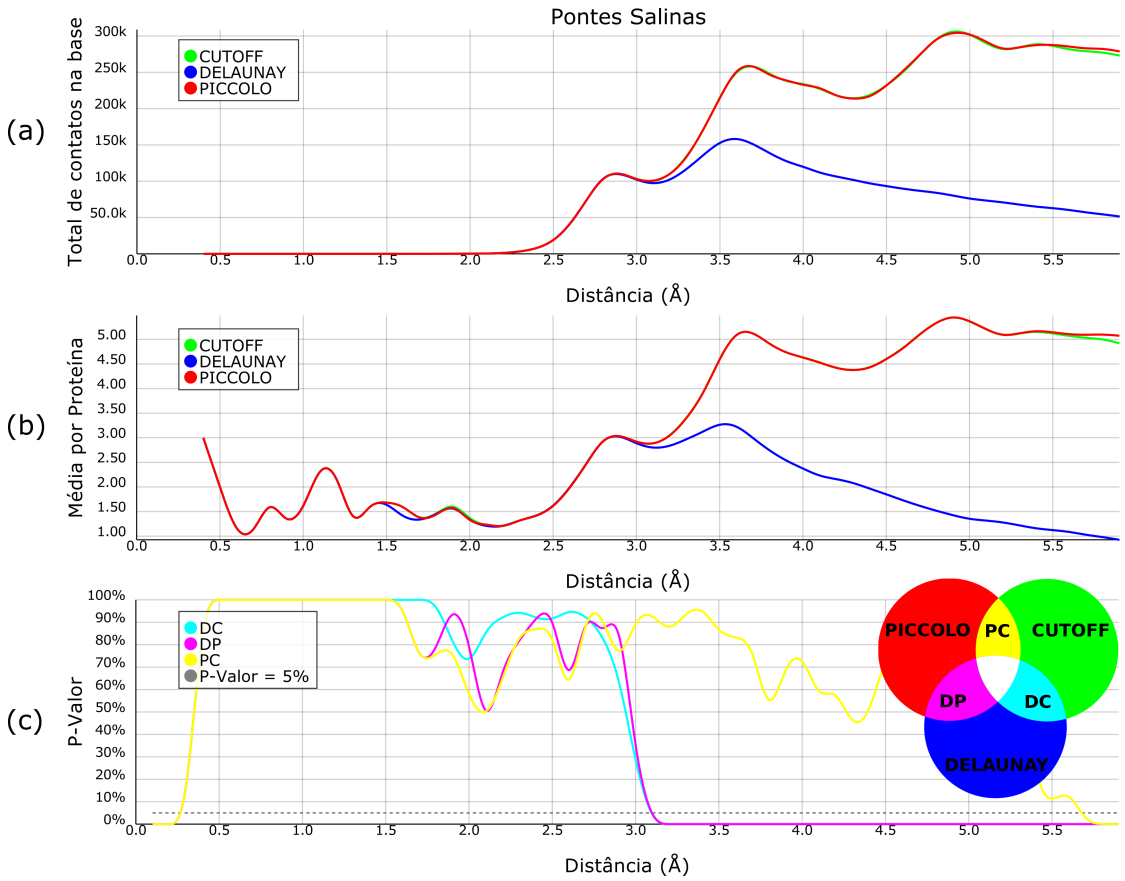


Figura 3.5: Análise comparativa dos paradigmas no cálculo de pontes salinas.

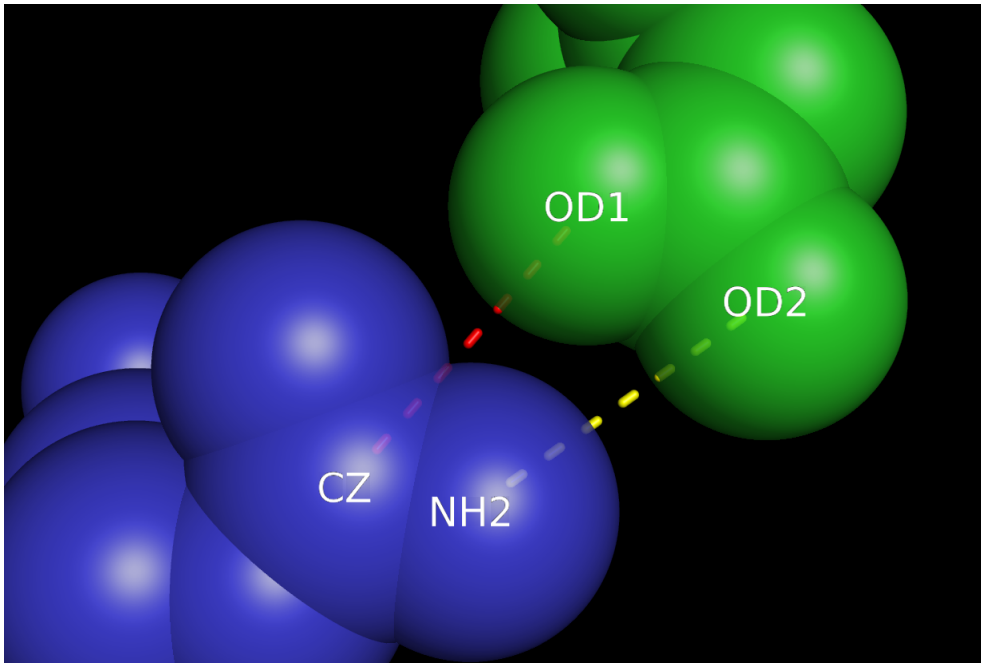


Figura 3.6: Exemplo de comparação de pontes salinas entre *delaunay* e *piccolo*. PDB 1A5G: ARG73:H e ASP55:I.

3.2.3.1 Exemplo

Para pontes salinas avaliamos um exemplo entre *piccolo* e *delaunay* à distância de 4,0 Å. A Figura 3.6 apresenta duas pontes salinas realizadas entre uma ARG (azul) e um ASP (verde). A ponte salina entre OD2 e NH2 (tracejado amarelo) ocorre sem nenhum empecilho para ambos paradigmas. Similar ao exemplo anterior das interações hidrofóbicas entre *cutoff* e *delaunay*, neste exemplo temos a impressão visual de que não há um oclusão, mas a tesselaçao de *delaunay* não permitiu a criação da aresta entre OD1 a CZ (tracejado vermelho), possivelmente porque as células criada no diagrama de Voronoi pelo OD1 e o NH2 interferiram no contato OD1-CZ , computando uma ponte salina apenas para o paradigma *piccolo*.

3.2.4 Empilhamentos aromáticos

Assim como na análise das ligações de hidrogênio, os empilhamentos aromáticos são outro tipo de contatos que depende fisicamente da angulação entre os anéis envolvidos podendo se apresentar em configurações *face-to-face*, *edge-to-edge* ou *edge-to-face*. O que se nota com essa análise é que claramente tanto angulação quanto oclusão são importantes evidências na eliminação de contatos espúrios. Note que a curva que ilustra o método baseado em *cutoff* diverge das outras em cerca de 3,6 Å tendo um crescimento enorme a partir desse limiar enquanto o método de *piccolo* que considera a angulação diverge em cerca de 3,7

Å mas seu crescimento é muito mais lento que o método puramente baseado em distância. Contudo o que se nota é que ainda assim, mesmo considerando os ângulos entre os anéis e estando essa adequada para o estabelecimento de um empilhamento aromático, o problema da oclusão pode fazer com que um grande percentual dos contatos encontrados sejam espúrio.

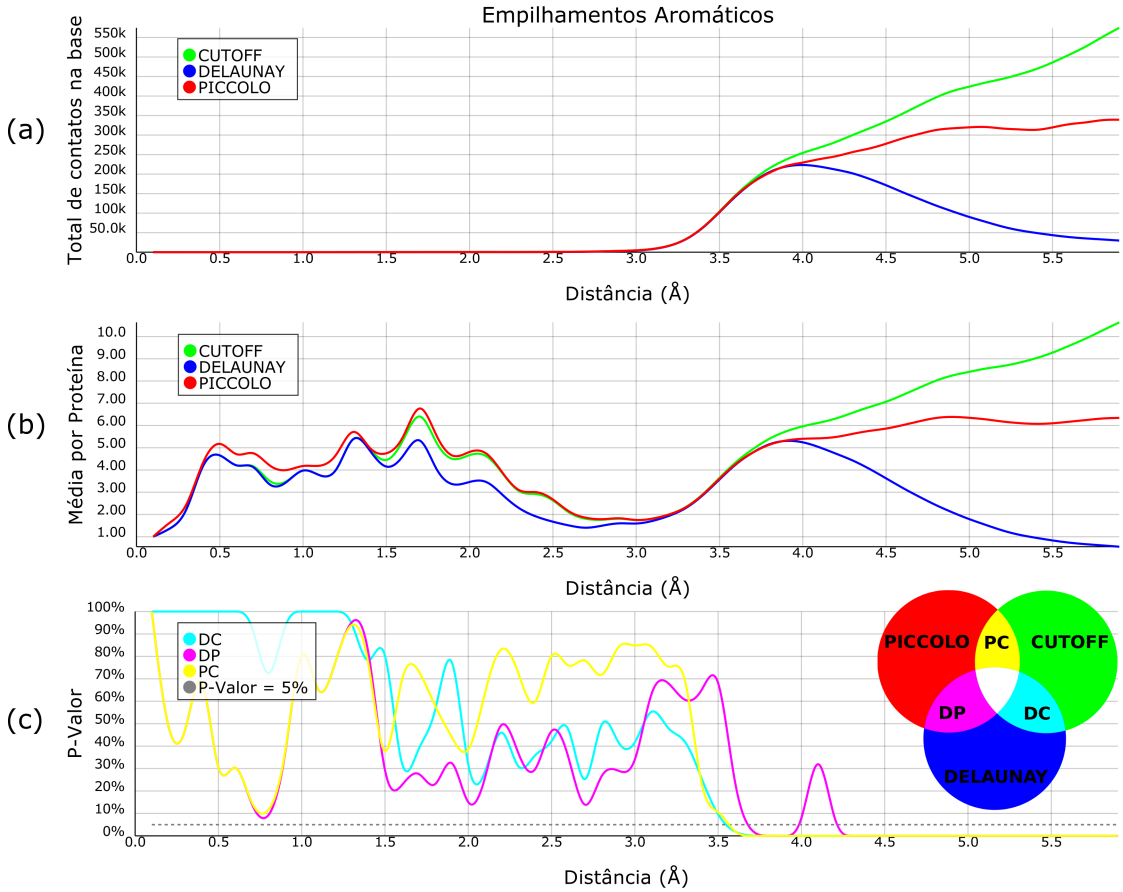


Figura 3.7: Análise comparativa dos paradigmas no cálculo de empilhamentos aromáticos.

É importante destacar que estamos aqui considerando que os contatos obtidos pelo método de *delaunay* são um subconjunto dos contatos obtidos pelo *piccolo*. Isso pode não ser verdade e isso não foi comprovado no presente trabalho sendo um trabalho futuro importante.

3.2.4.1 Exemplo

Para ilustrar o critério de angulação utilizado pelo *piccolo*, a Figura 3.8 mostra contatos entre anéis aromáticos comparando o paradigma *cutoff* e *piccolo*. Nela vemos dois contatos entre os anéis aromáticos de duas PHE (tracejado vermelho). Estes são válidos pelo paradigma *cutoff* (tracejado vermelho), que não utiliza angulação. O *piccolo* por sua vez não considera este exemplo como empilhamento aromático pois os angulos dos anéis não estão na configuração de (*face-to-face*, *edge-to-edge* ou *edge-to-face*).

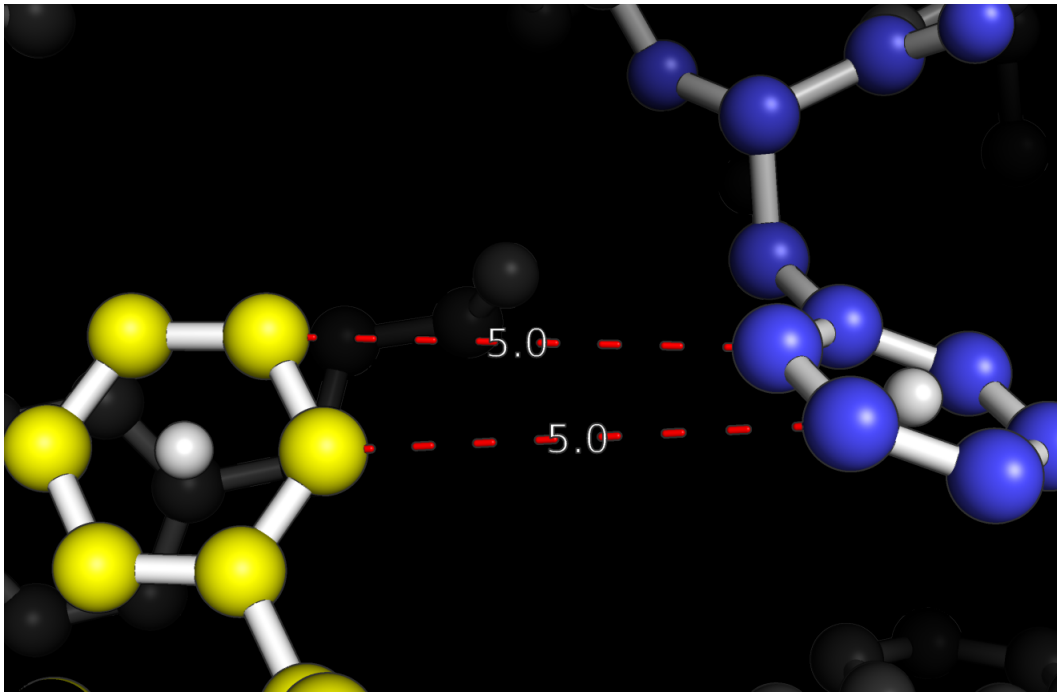


Figura 3.8: Exemplo de comparação de empilhamento aromático entre *cutoff* e *piccolo*. PDB 2LJY: PHE47:A e PHE47:B.

3.2.5 Resultados

A Figura 3.1c possui curvas que revelam até qual ponto de distância um par de paradigmas possam ser considerados iguais usando o teste de Wilcoxon. No início, próximo de 0,1 Å, a curva *piccolo-cutoff* (amarelo) indica que os paradigmas são iguais naquele ponto, mas não há certeza sobre isso, pois há poucos valores nesta região para afirmar uma hipótese de igualdade e com isso consideramos que em nenhum momento os pares *piccolo-cutoff* são iguais. A curva *delaunay-piccolo* (magenta) está sobreposta à curva *piccolo-cutoff* (amarelo) e por isso recebe a mesma análise. Quanto a curva *delaunay-cutoff* (ciano) temos uma queda na distância 3,6 Å indicando que até ali os paradigmas são iguais. Após esta análise chegamos a conclusão que o uso de critérios de angulação para ligações de hidrogênios afetam as comparação e com isso o paradigma *piccolo* é o mais aceitável para ligações de hidrogênio.

Quanto as interações hidrofóbicas, analisando a Figura 3.3c verifica-se que a curva *delaunay-cutoff* (ciano) oscila a partir de 1,5 Å, mas só atravessa o valor abaixo de 5% em 3,7 Å, o que faz sentido quando olhamos para a Figura 3.3a. Quanto aos pares com o paradigma *piccolo* (amarelo e magenta), percebe-se que as curvas muitas vezes ficam abaixo de 5%, porém esse acontecimentos oscilam e a curva *piccolo-cutoff* (amarelo) aparece pelo última vez acima de 5% em 3,5 Å assim como *delaunay-piccolo* (magenta). Com isso conclui-se que o paradigma *cutoff* se igualou em dois momentos de distância próximos, um em 3,7 Å

em par com o *delaunay* e 3,5 Å com o *piccolo*.

Pontes salinas criaram um perfis de curvas bem interessantes. Comparando a Figura 3.5a e 3.5c, vemos que elas são compatíveis. Em 3,0 Å, *delaunay-cutoff* (ciano) e *delaunay-piccolo* (magenta) se divergem da curva *piccolo-cutoff* (amarelo), o que era de se esperar quando a distância aumenta. Quando estamos em 5,6 Å, *piccolo-cutoff* (amarelo) se divergem pois começasse a se aproximar do raio de corte de 6,0 Å definido por *piccolo*.

Os perfis de curva de empilhamento aromático na Figura 3.7c são similares ao que se espera quando observa-se a Figura 3.7a. Apesar das muitas oscilações em todos os pares de paradigmas em nenhum momento as curvas ficam abaixo do limite de P-Valor até a distância de 3,5 Å. Esta é uma distância interessante, pois com exceção da curva *delaunay-piccolo* (magenta) que diverge a 3,6 Å, as curvas *delaunay-cutoff* (ciano) e *piccolo-cutoff* (amarelo) são consideradas distintas em 3,5 Å, ou seja, apenas um décimo de distância de diferença entre todas os pares de paradigmas. Porém há um pico em 4,1 Å na curva *delaunay-piccolo* (magenta) que consideramos ser um fator isolado na análise de empilhamento aromático por não condizer com a Figura 3.7a e por isso descartamos esta região. Conclui-se então que a distância de 3,5 Å é bem apropriada para empilhamentos aromáticos.

Em síntese à análise abordada a Tabela 3.1 contém as distâncias encontradas para cada par de paradigma por interação, assim bem como a média de distâncias.

	Distância de Divergência (Å)			
	DC	DP	PC	Média
Ligações de Hidrogênio	3,7	-	-	3,7
Interação Hidrofóbica	3,7	3,5	3,5	3,57
Ponte Salina	3,0	3,0	5,6	3,87
Empilhamento Aromático	3,5	3,6	3,5	3,53

Tabela 3.1: Distâncias de divergência de pares de paradigmas de contatos. DC = *delaunay-cuttof*; DP = *delaunay-piccolo*; PC = *piccolo-delaunay*.

Capítulo 4

Conclusão

Essa dissertação de mestrado teve como principal contribuição o projeto, implementação e disponibilização da base de dados CAPRI que é constituída por contatos inter-cadeia em interfaces proteína-proteína, deixando público para a comunidade científica todos os artefatos produzidos. Consideramos nessa base de dados as ligações de hidrogênio, interações hidrofóbicas, pontes salinas e empilhamentos aromáticos e comparamos três diferentes paradigmas baseados em diferentes tipos de informação que pode ser utilizada na prospecção de contatos: distância inter-atômica, oclusão e angulação. Outra importante contribuição do presente trabalho é a análise comparativa em termos do quantitativo de contatos prospectados com os diferentes paradigmas em diferentes limiares de distância e em nível atómico. Através desse trabalho pudemos concluir que apenas um limiar de distância adequado não garante que se obtenha apenas contatos legítimos mas que critérios de angulação são essenciais no especialmente no cálculo de ligações de hidrogênio.

Embora tenhamos conseguido realizar análises inéditas e obtido indícios a respeito da relevância de cada um dos critérios utilizados (aqui chamados paradigmas), esse trabalho na verdade mais abre espaço para outras discussões e levanta novas questões que serão investigadas em trabalhos futuros como por exemplo: Essas análises, resultados e conclusões se repetem em interações intra-cadeia? Poderiam esses resultados ser generalizados para interações proteína-ligante? Como calcular uma probabilidade de acerto na identificação de um contato com base apenas em um limiar de distância? Por que a curva de distribuição das pontes salinas é multimodal? O que cada modo representa? Poderiam esses modos ter alguma correlação com a esfericidade e/ou o raio de giro da proteína?

Acreditamos que essas questões poderão ser respondidas no futuro com a extensão da base de dados CAPRI com o acréscimo de interações intra-cadeia e proteína-ligante seguindo o mesmo esquema da base de dados com pequenas modificações bem como os programas e análises desenvolvidas podem ainda ser válidos nesses novos cenários de estudo.

Apêndice A

Tabela comparativa das propriedades fisico-químicas

		PICCOLO						SOBOLEV					
Resíduo	Átomo	Hidrofóbico	Aromático	Cátion	Aníon	Doador (Lig. de Hidrogênio)	Aceptor (Lig. de Hidrogênio)	Hidrofóbico	Aromático	Cátion	Aníon	Doador (Lig. de Hidrogênio)	Aceptor (Lig. de Hidrogênio)
ALA	C												
ALA	CA												
ALA	CB	X						X					
ALA	N					X						X	
ALA	O						X						X
ARG	C												
ARG	CA												
ARG	CB	x						X					
ARG	CD							S					
ARG	CG	x						X					
ARG	CZ			P				S					
ARG	N					X						X	
ARG	NE			P		X						X	

		PICCOLO						SOBOLEV					
Resíduo	Átomo	Hidrofóbico	Aromático	Cátion	Aníon	Doador (Lig. de Hidrogênio)	Aceptor (Lig. de Hidrogênio)	Hidrofóbico	Aromático	Cátion	Aníon	Doador (Lig. de Hidrogênio)	Aceptor (Lig. de Hidrogênio)
ARG	NH1			X		X				X		X	
ARG	NH2			X		X				X		X	
ARG	O						X						X
ASN	C												
ASN	CA												
ASN	CB	X						X					
ASN	CG							X					
ASN	N					X						X	
ASN	ND2					X						X	
ASN	O						X						X
ASN	OD1						X					S	X
ASP	C												
ASP	CA												
ASP	CB	X						X					
ASP	CG				P			S					
ASP	N					X						X	
ASP	O						X						X
ASP	OD1				X		X				X		X
ASP	OD2				X		X				X		X
CYS	C												
CYS	CA												
CYS	CB	X						X					
CYS	N					X						X	
CYS	O						X						X
CYS	SG					P	P						
GLN	C												
GLN	CA												

		PICCOLO						SOBOLEV					
Resíduo	Átomo	Hidrofóbico	Aromático	Cátion	Aníon	Doador (Lig. de Hidrogênio)	Aceptor (Lig. de Hidrogênio)	Hidrofóbico	Aromático	Cátion	Aníon	Doador (Lig. de Hidrogênio)	Aceptor (Lig. de Hidrogênio)
GLN	CB	X						X					
GLN	CD							S					
GLN	CG	X						X					
GLN	N					X						X	
GLN	NE2					P							
GLN	O						X						X
GLN	OE1						X						X
GLU	C												
GLU	CA												
GLU	CB	X						X					
GLU	CD				X			S					
GLU	CG	X						X					
GLU	N					X						X	
GLU	O						X						X
GLU	OE1				X		P				X		
GLU	OE2				X		X				X		X
GLY	C												
GLY	CA												
GLY	N					X						X	
GLY	O						X						X
HIS	C												
HIS	CA												
HIS	CB	X						X					
HIS	CD2		X	P				S	X				
HIS	CE1		X	P				S	X				
HIS	CG		X	P				S	X				
HIS	N					X						X	

		PICCOLO						SOBOLEV					
Resíduo	Átomo	Hidrofóbico	Aromático	Cátion	Aníon	Doador (Lig. de Hidrogênio)	Aceptor (Lig. de Hidrogênio)	Hidrofóbico	Aromático	Cátion	Aníon	Doador (Lig. de Hidrogênio)	Aceptor (Lig. de Hidrogênio)
HIS	ND1		X	X		X	P		X	X		X	
HIS	NE2		X	X		X	P		X	X		X	
HIS	O						X						X
ILE	C												
ILE	CA												
ILE	CB	X						X					
ILE	CD1	X						X					
ILE	CG1	X						X					
ILE	CG2	X						X					
ILE	N					X						X	
ILE	O						X						X
LEU	C												
LEU	CA												
LEU	CB	X						X					
LEU	CD1	X						X					
LEU	CD2	X						X					
LEU	CG	X						X					
LEU	N					X						X	
LEU	O						X						X
LYS	C												
LYS	CA												
LYS	CB	X						X					
LYS	CD	X						X					
LYS	CE							S					
LYS	CG	X						X					
LYS	N					X						X	
LYS	NZ			X		X				X		X	

		PICCOLO						SOBOLEV					
Resíduo	Átomo	Hidrofóbico	Aromático	Cátion	Aníon	Doador (Lig. de Hidrogênio)	Aceptor (Lig. de Hidrogênio)	Hidrofóbico	Aromático	Cátion	Aníon	Doador (Lig. de Hidrogênio)	Aceptor (Lig. de Hidrogênio)
LYS	O						X						X
MET	C												
MET	CA												
MET	CB	X						X					
MET	CE	X						X					
MET	CG	X						X					
MET	N					X						X	
MET	O						X						X
MET	SD	X					P	X					
PHE	C												
PHE	CA												
PHE	CB	X						X					
PHE	CD1	X	X					X	X				
PHE	CD2	X	X					X	X				
PHE	CE1	X	X					X	X				
PHE	CE2	X	X					X	X				
PHE	CG	X	X					X	X				
PHE	CZ	X	X					X	X				
PHE	N					X						X	
PHE	O						X						X
PRO	C												
PRO	CA												
PRO	CB	X						X					
PRO	CD							S					
PRO	CG	X						X					
PRO	N											S	
PRO	O						X						X

		PICCOLO						SOBOLEV					
Resíduo	Átomo	Hidrofóbico	Aromático	Cátion	Aníon	Doador (Lig. de Hidrogênio)	Aceptor (Lig. de Hidrogênio)	Hidrofóbico	Aromático	Cátion	Aníon	Doador (Lig. de Hidrogênio)	Aceptor (Lig. de Hidrogênio)
SER	C												
SER	CA												
SER	CB							S					
SER	N					X						X	
SER	O						X						X
SER	OG					X	P					X	
THR	C												
THR	CA												
THR	CB							S					
THR	CG2	X						X					
THR	N					X						X	
THR	O						X						X
THR	OG1					X	P					X	
TRP	C												
TRP	CA												
TRP	CB	X						X					
TRP	CD1		X					S	X				
TRP	CD2	X	X					X	X				
TRP	CE2		X					S	X				
TRP	CE3	X	X					X	X				
TRP	CG	X	X					X	X				
TRP	CH2	X	X					X	X				
TRP	CZ2	X	X					X	X				
TRP	CZ3	X	X					X	X				
TRP	N					X						X	
TRP	NE1		X			X			X			X	
TRP	O						X						X

		PICCOLO						SOBOLEV					
Resíduo	Átomo	Hidrofóbico	Aromático	Cátion	Aníon	Doador (Lig. de Hidrogênio)	Aceptor (Lig. de Hidrogênio)	Hidrofóbico	Aromático	Cátion	Aníon	Doador (Lig. de Hidrogênio)	Aceptor (Lig. de Hidrogênio)
TYR	C												
TYR	CA												
TYR	CB	X						X					
TYR	CD1	X	X					X	X				
TYR	CD2	X	X					X	X				
TYR	CE1	X	X					X	X				
TYR	CE2	X	X					X	X				
TYR	CG	X	X					X	X				
TYR	CZ		X					S	X				
TYR	N					X						X	
TYR	O						X						X
TYR	OH					X						X	
VAL	C												
VAL	CA												
VAL	CB	X						X					
VAL	CG1	X						X					
VAL	CG2	X						X					
VAL	N					X						X	
VAL	O						X						X

Tabela A.1: Tabela comparativa das propriedades fisico-químicas quanto as definições de PICCOLO e SOBOLEV. X: ambos definições são iguais; P: definido somente por PICCOLO; S: definido somente por SOBOLEV

Referências Bibliográficas

Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N. & Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*, 28(1):235–242.

Bernal, J. D. & Crowfoot, D. (1934). X-ray photographs of crystalline pepsin. *Nature*, 133(3369):794–795.

Bickerton, G. R.; Higueruelo, A. P. & Blundell, T. L. (2011). Comprehensive, atomic-level characterization of structurally characterized protein-protein interactions: the piccolo database. *BMC bioinformatics*, 12(1):313.

Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M.-C.; Estreicher, A.; Gasteiger, E.; Martin, M. J.; Michoud, K.; O'Donovan, C.; Phan, I. et al. (2003). The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic acids research*, 31(1):365–370.

Bostock, M.; Ogievetsky, V. & Heer, J. (2011). D 3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309.

Bowie, J. U.; Luthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253(5016):164–170.

Davies, M. N.; Toseland, C. P.; Moss, D. S. & Flower, D. R. (2006). Benchmarking pka prediction. *BMC biochemistry*, 7(1):18.

de Melo, R.; Ribeiro, C.; Murray, C.; Veloso, C.; da Silveira, C.; Neshich, G.; Meira Jr, W.; Carceroni, R. & Santoro, M. (2007). Finding protein-protein interaction patterns by contact map matching. *Genet. Mol. Res*, 6(4):946–963.

Delano, W. L. (2002). The PyMOL Molecular Graphics System.

Fassio, A. V. (2015). napoli: uma ferramenta web para análise de interações proteína-ligante. Dissertação de mestrado, Universidade Federal de Minas Gerais, Universidade Federal de Minas Gerais, Belo Horizonte.

Fogolari, F.; Brigo, A. & Molinari, H. (2002). The poisson-boltzmann equation for biomolecular electrostatics: a tool for structural biology. *Journal of Molecular Recognition*, 15(6):377–92.

Franceschini, A.; Szklarczyk, D.; Frankild, S.; Kuhn, M.; Simonovic, M.; Roth, A.; Lin, J.; Minguez, P.; Bork, P.; von Mering, C. et al. (2013). String v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, 41(D1):D808–D815.

Geoffrey M. Cooper, R. E. H. (2006). *The Cell: A Molecular Approach*. Sinauer Associates.

Godzik, A.; Kolinski, A. & Skolnick, J. (1992). Topology fingerprint approach to the inverse protein folding problem. *Journal of molecular biology*, 227(1):227–238.

Gonçalves-Almeida, V.; Pires, D. E.; de Melo-Minardi, R. C.; da Silveira, C. H.; Meira, W. & Santoro, M. M. (2012). Hydropace: understanding and predicting cross-inhibition in serine proteases through hydrophobic patch centroids. *Bioinformatics*, 28(3):342–349.

Hamelryck, T. & Manderick, B. (2003). Pdb file parser and structure class implemented in python. *Bioinformatics*, 19(17):2308–2310.

Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *Journal of molecular biology*, 233(1):123–138.

Hubbard, S. J. & Thornton, J. M. (1993). Naccess. *Computer Program, Department of Biochemistry and Molecular Biology, University College London*, 2(1).

Jones, E.; Oliphant, T.; Peterson, P. et al. (2001). SciPy: Open source scientific tools for Python. [Online; accessed 2015-07-04].

Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R.; Wyckoff, H. & Phillips, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662–666.

Kessel, A. & Ben-Tal, N. (2010). *Introduction to proteins: structure, function, and motion*. CRC Press.

Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *Journal of molecular biology*, 55(3):379–IN4.

Leinonen, R.; Akhtar, R.; Birney, E.; Bower, L.; Cerdeno-Tárraga, A.; Cheng, Y.; Cleland, I.; Faruque, N.; Goodgame, N.; Gibson, R. et al. (2010). The european nucleotide archive. *Nucleic acids research*.

Lesk, A. M. & Andrade, A. E. (2008). *Introdução à bioinformática*. Artmed.

Manavalan, P. & Ponnuswamy, P. (1977). A study of the preferred environment of amino acid residues in globular proteins. *Archives of biochemistry and biophysics*, 184(2):476–487.

Mancini, A. L.; Higa, R. H.; Oliveira, A.; Dominiquini, F.; Kuser, P. R.; Yamagishi, M. E. B.; Togawa, R. C. & Neshich, G. (2004). Sting contacts: a web-based application for identification and analysis of amino acid contacts within protein structure and across protein interfaces. *Bioinformatics*, 20(13):2145–2147.

McDonald, I.; Naylor, D.; Jones, D. & Thornton, J. (1993). Hbplus computer program. *Department of Biochemistry and Molecular Biology, University College, London, UK*.

Miyazawa, S. & Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, 18(3):534–552.

Murzin, A. G.; Brenner, S. E.; Hubbard, T. & Chothia, C. (1995). Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, 247(4):536–540.

Nelson, D. L. & Cox, M. M. (2014). *Princípios de Bioquímica de Lehninger*. Porto Alegre: Artmed, 6 edição.

Plaxco, K. W.; Simons, K. T. & Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *Journal of molecular biology*, 277(4):985–994.

Richards, F. M. (1974). The interpretation of protein structures: total volume, group volume distributions and packing density. *Journal of molecular biology*, 82(1):1–14.

Samudrala, R. & Moult, J. (1998). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *Journal of molecular biology*, 275(5):895–916.

Sanger, F. (1988). Sequences, sequences, and sequences. *Annual review of biochemistry*, 57(1):1–29.

Silveira, C. H.; Pires, D. E.; Minardi, R. C.; Ribeiro, C.; Veloso, C. J.; Lopes, J. C.; Meira, W.; Neshich, G.; Ramos, C. H.; Habesch, R. et al. (2009). Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, 74(3):727–743.

Sobolev, V.; Sorokine, A.; Prilusky, J.; Abola, E. E. & Edelman, M. (1999). Automated analysis of interatomic contacts in proteins. *Bioinformatics*, 15(4):327–332.

Stryer, L.; Tymoczko, J. L. & Berg, J. M. (2004). *Bioquímica*. Guanabara Koogan.

Tsai, C.-J.; Lin, S. L.; Wolfson, H. J. & Nussinov, R. (1997). Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein science: a publication of the Protein Society*, 6(1):53.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, pp. 80–83.