ALEXANDRE VICTOR FASSIO

nAPOLI: UMA FERRAMENTA WEB PARA ANÁLISE DE INTERAÇÕES PROTEÍNA-LIGANTE

Belo Horizonte, MG

Junho de 2015

ALEXANDRE VICTOR FASSIO

nAPOLI: UMA FERRAMENTA WEB PARA ANÁLISE DE INTERAÇÕES PROTEÍNA-LIGANTE

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: Raquel Cardoso de Melo Minardi
Coorientador: Sabrina de Azevedo Silveira

Belo Horizonte, MG

Junho de 2015

043 Fassio, Alexandre Victor.
nAPOLI: uma ferramenta Web para análise de interações proteína-ligante [manuscrito] / Alexandre Victor Fassio. – 2015.
131 f. : il. ; 29,5 cm.

Orientadora: Raquel Cardoso de Melo Minardi. Co-orientadora: Sabrina de Azevedo Silveira.
Dissertação (mestrado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas.

1. Reconhecimento molecular - Teses. 2. Proteína de ligação - Teses. 3. Visualização - Teses. 4. Interfaces visuais. 5. Interações. 6. Quinase 2 dependente de ciclina - Teses. 7. Bioinformática. I. Minardi, Raquel Cardoso de Melo. II. Silveira, Sabrina de Azevedo. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática da UFMG
Avenida Presidente Antônio Carlos, 6627 – Pampulha
31270-901 - Belo Horizonte – MG
Endereço eletrônico: bioinfo@icb.ufmg.br 55 31 3409-2554

"nAPOLI: Uma ferramenta web para análise de interações proteína-ligante"

Alexandre Victor Fassio

Dissertação aprovada pela banca examinadora constituída pelos Professores:

Profa Raquel Cardoso de Melo Minardi - Orientadora
UFMG

Profa Sabrina de Azevedo Silveira - Co-Orientadora
UFMG

Profa Rafaela Salgado Ferreira
UFMG

Profa Valdete Maria Gonçalves Almeida
UNIBH

Belo Horizonte, 30 de junho de 2015.

Dr. Vasco Ariston de C. Azevedo
Prof. Titular e Coordenador do Programa de Pós Graduação em Bioinformática
ICB/UFMG

Agradecimentos

Primeiramente agradeço a Deus por tudo que alcancei, por ser quem sou, pela minha saúde, perseverança e inteligência.

Aos meus pais, pois sem eles nada disso seria possível. Meus pais que sempre nos incentivaram ao estudo, que nos deram condições para que isso fosse possível, foram professores da vida ao nos ensinar boas virtudes, mostrando-nos o caminho certo a ser seguido. Agradeço também à minha irmã e meu irmão por estarem sempre ao meu lado, pela amizade e apoio.

À minha preciosa, amada e amiga, Luanna. Por sempre estar me apoiando e me incentivando mesmo nos altos e baixos. Sei que a distância fez-nos crescer ainda mais. A saudade que esteve ao nosso lado a maior parte do tempo sempre foi recompensada e sei que nos fortalece cada dia mais. Nossa paciência rendeu frutos. Sei também que essa foi apenas a primeira etapa de muitas que virão e sei que estaremos sempre juntos ao longo desta e outras vidas.

Agradeço às amizades que foram estabelecidas nesta etapa de minha vida. Agradeço aos amigos Igor e Thiago pela convivência e amizade.

Agradeço a todos os amigos da Bioinformática e amigos do Laboratório de Bioinformática e Sistemas (LBS) pela amizade e contribuição que tiveram em meu estudo, trabalho e formação. Agradeço ao Laerte por ter me ajudado diversas vezes e por estar sempre de prontidão para tirar minhas dúvidas.

Agradeço à Raquel pela oportunidade de realizar este trabalho, por contribuir diretamente para a minha formação e pelo que aprendi durante todo o período do mestrado. Por todo apoio e incentivo. Obrigado por estar sempre de prontidão respondendo emails nas férias, no fim de semana e qualquer dia que fosse.

Agradeço à Sabrina pela amizade, pela minha formação e por ajudar-me sempre que precisei. Foram muitos trabalhos realizados durante esse período com sua colaboração.

Agradeço ainda a todos professores da Universidade Federal de Minas Gerais que durante estes 2 anos contribuíram para a minha formação e conhecimento.

Resumo

Elucidar os mecanismos envolvidos no reconhecimento molecular e quais forças contribuem para o reconhecimento é um problema central na biologia, uma vez as interações entre duas moléculas são de extrema importância para os sistemas biológicos como um todo e são difíceis de serem preditos até mesmo para pequenas moléculas. Portanto, propomos uma ferramenta que visa ser fácil e intuitiva a partir do uso de estratégias visuais para descrever os padrões e tipos de interações estabelecidas entre proteínas e seus ligantes. Para tanto, propomos nAPOLI (Analysis of PrOtein Ligand Interactions), uma ferramenta *web* interativa para o estudo em larga escala das interações proteína-ligante utilizando estratégias visuais e análises estatísticas. Realizamos um estudo de caso que consistiu de 73 inibidores para a proteína humana CDK2 (Ciclin-dependent kinase 2). nAPOLI mostrou ser muito útil, pois ajudou na descoberta do que era conservado em um conjunto de ligantes e quais interações foram estabelecidas com o receptor. Através da ferramenta, fomos capazes de confirmar resultados experimentais da literatura. Além disso, apesar da existência de outras ferramentas para analisar interações proteína-ligante, nenhuma delas apresenta análises estatísticas, visuais e interativas em larga escala. A ferramenta pode ser acessada em: http://www.napoli.dcc.ufmg.br/

Palavras-chave: Interações, Contatos, Reconhecimento molecular, Ligante, Receptor, CDK2, Proteína, Visualização de dados, Interfaces visuais, Interativo.

Abstract

Elucidating the mechanisms involved in the molecular recognition and which forces contribute to the recognition is a central problem in biology, since the interactions between two molecules are extremely important to biological systems as a whole and are very toilsome to be predicted even for small molecules. Therefore, we propose an easy and intuitive tool through visual strategies to depict the patterns and the types of interactions established between proteins and their ligands. In order to achieve it, we propose nAPOLI (Analysis of PrOtein Ligand Interactions), an interactive web tool to study the protein-ligand interactions in large scale by using visual strategies and statistical analysis. We performed a case study consisting of 73 inhibitors to the human protein CDK2 (Ciclin-dependent kinase 2). nAPOLI showed to be very useful as it shed some light on the problem of discovering what is conserved in a set of ligands and which interactions they establish with a receptor. Through this tool, we were able to confirm literature experimental results. Despite of the existance of other tools to analyse protein-ligand interactions, none of them present statistical, visual and interactive analysis in large scale. The tool can be accessed in: http://www.napoli.dcc.ufmg.br/

Keywords: Interactions, Contacts, Molecular recognition, Ligand, Receptor, CDK2, Protein, Information visualization, Visual interfaces, Interactive.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

1.1 Proteínas

Grande parte do funcionamento e dos processos biológicos em organismos vivos depende de macromoléculas conhecidas como proteínas. Essas macromoléculas são consideradas as mais abundantes, sendo encontradas em todas as células de um organismo. Dessa forma, as proteínas são consideradas vitais para os seres vivos em geral devido às diversas funções desempenhadas [Kessel e Ben-Tal, 2010; Nelson e Cox, 2014].

Não somente existe uma diversidade de proteínas como também as funções desempenhadas por elas são bastante variadas [Kessel e Ben-Tal, 2010; Nelson e Cox, 2014; Stryer et al., 2004; Zvelebil e Baum, 2008]. Proteínas como a hemoglobina, por exemplo, possuem um papel importante no transporte de oxigênio na corrente sanguínea de vertebrados e de alguns invertebrados; a imunoglobulina é importante, pois é um anticorpo que faz parte do sistema imunológico, atuando em defesa do organismo contra agentes externos que podem atrapalhar o funcionamento pleno de processos biológicos e acarretar doenças nos organismos; por outro lado, algumas proteínas atuam como enzimas e desempenham um papel importante na catálise de reações químicas e no metabolismo dos organismos [Dunn, 2010; Kessel e Ben-Tal, 2010].

Em geral, a função desempenhada por uma proteína depende de sua estrutura tridimensional que é constituída a partir de um repertório de moléculas, conhecidas como aminoácidos. A gama de diversidade funcional e especificidades encontradas atualmente nas proteínas originou-se ao longo de milênios devido às pressões seletivas do ambiente; isto desencadeou o processo evolutivo que culminou na origem de diversas proteínas e na ampla extensão de funções desempenhadas [Branden e Tooze, 1999].

1.1.1 Estrutura de proteínas

As estruturas tridimensionais de proteínas, como mencionado anteriormente, dependem de uma sequência de moléculas chamadas aminoácidos. Esses aminoácidos são organizados de forma sequencial e linear, formando, assim, uma cadeia de polipeptídeos comumente chamada de estrutura primária. Os 20 aminoácidos mais comumente encontrados nos seres vivos são apresentados na Tabela 1.1.

Tabela 1.1. Aminoácidos mais comumente encontrados nos seres vivos e suas abreviações com 3 e 1 letra.

Aminoácido	Símbolo de 3 letras	Abreviação
Alanina	ALA	A
Arginina	ARG	R
Asparagina	ASN	N
Aspartato ou Ácido aspártico	ASP	D
Cisteína	CYS	C
Fenilalanina	PHE	F
Glicina	GLY	G
Glutamato ou Ácido glutâmico	GLU	E
Glutamina	GLN	Q
Histidina	HIS	H
Isoleucina	ILE	I
Leucina	LEU	L
Lisina	LYS	Y
Metionina	MET	M
Prolina	PRO	P
Serina	SER	S
Tirosina	TYR	Y
Treonina	THR	T
Triptofano	TRP	W
Valina	VAL	V

Todos os aminoácidos possuem uma estrutura básica comum que é denominada cadeia principal. Nela, estão presentes os grupos funcionais carboxila e amina. Entretanto, o que possibilita as características e propriedades químicas específicas de cada aminoácido é sua cadeia lateral (grupo R na Figura 1.1). Assim, cada aminoácido pode ser classificado de acordo com suas propriedades físico-químicas, tais como, cargas, hidrofobicidade, polaridade, etc.

Como mencionado anteriormente, uma cadeia de polipeptídeos envolve um conjunto de aminoácidos dispostos sequencialmente e interconectados por ligações peptídicas formadas através de uma reação entre o grupo funcional carboxila de um aminoácido

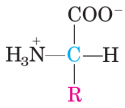


Figura 1.1. A estrutura básica de um aminoácido: uma cadeia principal e uma cadeia lateral (grupo R) [Nelson e Cox, 2014].

e o grupo funcional amina do aminoácido seguinte. Esse tipo de ligação envolve um tipo de ligação química denominada interação covalente em que há um compartilhamento de elétrons entre os átomos envolvidos na ligação.

Quando uma ligação peptídica é estabelecida, há a liberação de uma molécula de água, que é composta a partir da hidroxila (OH) do grupo carboxila de um aminoácido e de um hidrogênio (H) do grupo amina de outro aminoácido (Figura 1.2). Assim, estes aminoácidos passam a ser chamados de resíduos de aminoácido ou apenas resíduos.

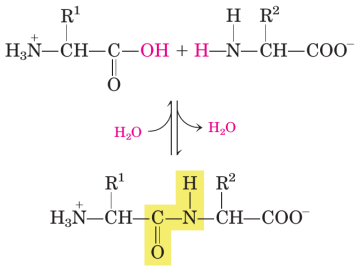


Figura 1.2. Formação de uma ligação peptídica entre o grupo funcional carboxila (aminoácido com radical R1) e o grupo amina (aminoácido com radical R2) com a liberação de uma molécula de água (H_2O) [Nelson e Cox, 2014].

Assim, são as propriedades intrínsecas e a forma na qual a sequência de aminoácidos é disposta que influenciam diretamente no enovelamento de uma proteína e sua correspondente função. Tanto é verdade que uma simples troca em um aminoácido da sequência, pode produzir efeitos importantes nos organismos, como é o caso da anemia falciforme causada pela substituição de um ácido glutâmico por uma valina [Branden e Tooze, 1999; Nelson e Cox, 2014].

O enovelamento é o processo pelo qual uma proteína atinge sua conformação nativa e funcional e é um fenômeno ainda não completamente compreendido pela ciência dada sua complexidade. Em termos simplificados, um enovelamento pode ser considerado como o processo pelo qual a proteína dobra-se e enrola-se em torno de si mesma e adquire sua conformação estrutural [Nelson e Cox, 2014].

A estrutura de uma proteína ainda pode ser classificada em mais três níveis estruturais (Figura 1.3):

- Estrutura secundária: é o primeiro arranjo estrutural tomado pela proteína no processo de enovelamento. Basicamente, refere-se ao arranjo espacial dos resíduos ao longo de um segmento da cadeia polipeptídica e que são estabilizadas por pontes de hidrogênio entre os resíduos próximos no segmento. As α-hélices e folhas-β são dois exemplos de tais estruturas.

- Estrutura terciária: formado pelo arranjo tridimensional das estruturas secundárias numa cadeia polipeptídica. São estabilizadas por ligações não-covalentes como pontes salinas, interações hidrofóbicas e pontes de hidrogênio ou mesmo por pontes dissulfeto que são classificadas como ligações covalentes. Tais interações podem ocorrer entre resíduos localizados em posições distantes da cadeia polipeptídica, diferentemente do que ocorre na estrutura secundária que é mais local.

- Estrutura quaternária: formada a partir de um arranjo tridimensional de mais de uma estrutura terciária, ou seja, consiste em complexos nos quais duas ou mais cadeias proteicas interagem.

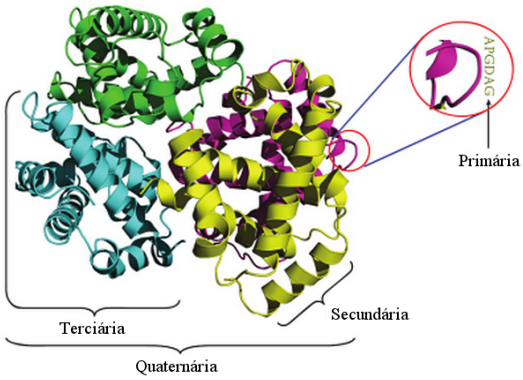


Figura 1.3. Os quatro níveis da estrutura de uma proteína. Adaptado de Kessel e Ben-Tal [2010].

São majoritariamente essas estruturas que diferenciam uma proteína em termos funcionais, sendo que algumas proteínas atuam como hormônios, outras como enzimas ou mesmo como anticorpos. Assim, cada proteína possui características únicas de acordo com sua estrutura primária e sua estrutura terciária final.

Além disso, muitas das funções desempenhadas por uma proteína e, consequentemente, muitos dos processos biológicos relacionados a elas dependem também de interações altamente específicas que ocorrem entre duas ou mais moléculas, como no caso das enzimas, por exemplo. Dessa forma, muitos estudos são realizados para tentar

compreender o princípio de reconhecimento e interação de uma proteína com diversas outras moléculas e macromoléculas. Há pesquisas que visam entender os padrões existentes entre as várias interações que ocorrem entre uma proteína e moléculas (ligantes); entre duas proteínas; ou ainda entre proteínas e DNA ou RNA. Neste trabalho, o foco será investigar as interações de uma proteína e ligantes (não proteicos, ou pequenas moléculas), isto é, interações proteína-ligante.

1.2 Reconhecimento molecular

O termo reconhecimento molecular se refere às interações específicas que ocorrem entre duas ou mais moléculas e envolvem interações inter-moleculares.

Dessa forma, é necessário que haja uma complementariedade química e geométrica entre as partes envolvidas na interação. Diversos são os papéis biológicos desempenhados por esse tipo de interação, podendo-se citar as interações que ocorrem entre hormônios e receptores; substratos e enzimas; antígenos e anticorpos; transportadores e solutos; ácidos nucleicos e ribossomos; entre outros [Dunn, 2010; Kessel e Ben-Tal, 2010].

Portanto, o reconhecimento molecular entre duas ou mais moléculas é de extrema importância para os sistemas e é um processo controlado por interações não-covalentes específicas [Böhm, 2005; Gellman, 1997]. Ainda, segundo [Dunn, 2010], interações entre proteína e ligante são fundamentais para quase todos os processos que ocorrem nos organismos vivos.

As diversas funções desempenhadas por proteínas são, em grande parte, dependentes de sítios específicos de ligação para outras moléculas. Essas moléculas se ligam à proteína receptora através de uma complementariedade molecular que geralmente ocorre de forma a obter uma conformação que seja energeticamente favorável [Böhm, 2005].

A conformação de uma proteína se refere aos arranjos espaciais de seus átomos constituintes. Quando uma ligação entre uma proteína e um ligante ocorre, muitas vezes, a conformação da proteína é alterada para que a ligação seja a mais favorável possível. Considera-se uma conformação favorável energeticamente aquela com a menor energia livre e com o número máximo de interações fracas. Essas interações são importantes para a estabilização de uma proteína em seu estado enovelado e oferecem a especificidade necessária para o reconhecimento molecular. Além disso, uma conformação somente será passível de ser obtida quando as interações covalentes (ligações peptídicas) entre os aminoácidos de uma proteína não forem degradadas [Nelson e Cox,

2014].

De fato, tal como explica Koshland [1958] com seu modelo de encaixe induzido, tanto o ligante quanto seu receptor são estruturas flexíveis e que, portanto, possuem a capacidade de alterar sua conformação [Boehr e Wright, 2008; Caceres et al., 2008; Csermely et al., 2010; Nelson e Cox, 2014]. Nesse caso, o ligante induz um comportamento de mudança conformacional no receptor a fim de aumentar a complementariedade molecular entre ambos. Por outro lado, o modelo de encaixe selecionado [Tsai et al., 1999, 2001] propõe que o ligante seleciona dentre os vários estados conformacionais de uma proteína aquele mais adequado à complementariedade das duas moléculas [Boehr e Wright, 2008; Caceres et al., 2008; Csermely et al., 2010]. Um modelo não necessariamente refuta o outro. Segundo Berger et al. [1999]; Boehr e Wright [2008]; Dunn [2010]; Foote e Milstein [1994]; James et al. [2003]; Kessel e Ben-Tal [2010], ambos modelos podem coexistir em certos sistemas.

Dessa forma, diversos modelos foram sendo propostos ao longo dos anos a fim de trazer alguma luz ou mesmo tentar encontrar respostas a uma pergunta intrigante: como uma proteína pode se ligar tão especificamente e seletivamente à apenas certas moléculas em um ambiente repleto de outras moléculas?

Diversos estudos focam em entender como ocorre e quais as forças envolvidas no processo de reconhecimento molecular. Segundo Grunenberg [2011], essa área ainda representa um desafio computacional mesmo para sistemas biológicos mais simples, uma vez que diversos eventos e processos estão envolvidos no reconhecimento molecular. Desse modo, a consideração de todos eventos e processos fazem com que o processamento computacional seja extremamente demorado.

1.2.1 O que é um ligante?

Um ligante é uma molécula que se liga a outra (receptor) para desempenhar uma função biológica específica [Andrabi et al., 2009]. Ligantes podem ser dos mais diversos tipos desde íons, carboidratos tais como glicose e galactose, RNAs, DNAs e até mesmo outras proteínas.

De forma geral, todos esses tipos de interações são chamadas de interações proteína-ligante. Porém, definições mais específicas podem ser utilizadas para descrever cada uma dessas interações. Por exemplo, interações entre duas proteínas são frequentemente chamadas de interações proteína-proteína; RNAs e DNAs interagindo com proteínas descrevem as interações RNA-proteína e DNA-proteína, respectivamente. Segundo [Andrabi et al., 2009], em química, o termo ligante é usado para descrever átomos ou grupos de átomos (pequenas moléculas) que se ligam à outras

moléculas. Para tanto, neste trabalho o termo interação proteína-ligante será utilizado como referência às interações que ocorrem entre uma proteína e um átomo – íons, por exemplo – ou pequenas moléculas – glicose, por exemplo.

Entretanto, independentemente do tipo de interação, uma ligação entre um ligante e seu receptor se dá por interações não-covalentes e algumas vezes por interações covalentes. A importância das interações não-covalentes se deve ao fato de que as forças envolvidas nesse processo são de caráter fraco e, portanto, possuem a característica de serem reversíveis (ligante e proteína ficam livres para outras interações). Segundo Nelson e Cox [2014], a natureza transitória das interações proteína-ligante é fundamental para a vida, pois permite que um organismo responda de maneira rápida e reversível a mudanças ambientais e condições metabólicas.

Os papéis desempenhados por ligantes em processos biológicos são bastante diversos e incluem: subtratos em reações catalíticas; a regulação de atividades metabólicas e de transmissão de sinais; redes de comunicação celular tal como ocorre com a transmissão de sinais via hormônios; grupos prostéticos que ajudam as proteínas a desempenharem sua função – um exemplo é a hemoglobina e seu grupo heme responsável por ligar e transportar oxigênio no sangue; defesa e ataque, isto é, certos ligantes são utilizados como substâncias tóxicas na defesa contra outros organismos ou como forma de ataque [Kessel e Ben-Tal, 2010].

Todos esses papéis são desempenhados graças às interações que ocorrem em regiões específicas da proteína receptora, que são conhecidas como sítio de ligação. Esses sítios devem apresentar a complementaridade geométrica e química necessária para que ocorra o reconhecimento molecular e a ligação entre ambas moléculas [Lahti et al., 2012]. Em enzimas, esse sítio de ligação no qual o ligante interage também é conhecido como sítio ativo [Nelson e Cox, 2014].

Já em algumas proteínas, as funções desempenhadas dependem de mais de um sítio de ligação. Segundo Dunn [2010], a ligação de uma molécula em um dado sítio da proteína pode influenciar completamente outro sítio de ligação mesmo que eles estejam distantes entre si. A influência pode muitas vezes refletir em uma alteração conformacional e, consequentemente, modificar a atividade do outro sítio [Mackinnon et al., 2014]. Esse efeito é conhecido como "efeito alostérico" e pode tanto ativar ou inibir a ligação de uma molécula.

A maneira como duas ou mais moléculas são capazes de se reconhecerem de forma específica mesmo em um ambiente com diversas outras moléculas disponíveis se apresenta como um problema intrigante e tem demandado estudos.

Portanto, perguntas como 'como é possível uma proteína reconhecer um ou mais ligantes de forma específica e eficiente?' ou 'como é possível um ligante interagir com

diversas proteínas diferentes?' são questões que são alvos de pesquisa, pois permitem entender a dinâmica, o funcionamento e como se dá o reconhecimento molecular entre duas ou mais moléculas. Para tanto, ferramentas computacionais que auxiliem no estudo desse tipo de processo se mostram ser promissoras dada a complexidade do problema e o volume e complexidade dos dados disponíveis para tais estudos.

1.3 Interações inter-moleculares

A especificidade na forma como as moléculas se reconhecem se deve a uma complementariedade geométrica e química. Essa última pode ser descrita pelas forças eletrostáticas que ocorrem de acordo com os átomos presentes nas moléculas, podendo estes estarem carregados positivamente ou negativamente ou mesmo serem neutros (polares ou apolares). Essas interações se estabelecem por forças não covalentes que são consideradas fracas. Nesse grupo enquadram-se as interações de van der Waals, pontes de hidrogênio, empilhamento aromático, interação hidrofóbica e interações eletrostáticas como as forças repulsivas e atrativas (ponte salina) [Andrabi et al., 2009; Pickett, 2005; Caceres et al., 2008; Nelson e Cox, 2014].

As interações de van der Waals podem ser explicadas pela distribuição da carga eletrônica de um átomo em um dado instante. Como os elétrons estão em constante movimento em torno do núcleo de um átomo, isto faz com que a distribuição de cargas seja assimétrica e forme um polo momentâneo, porém, suficiente para induzir cargas opostas em um outro átomo (polarizar um outro átomo). Dessa forma, os dois átomos são atraídos entre si até uma distância mínima, pois, a partir dessa distância haverá a sobreposição e repulsão entre os elétrons destes átomos tal como descrito pelo princípio de exclusão de Pauli (dois elétrons não podem ocupar o mesmo espaço em um mesmo instante). A distância mínima depende dos raios de van der Waals de cada átomo, que pode ser definida como a distância ótima de seu núcleo até a camada eletrônica mais externa de um átomo vizinho [Kessel e Ben-Tal, 2010; Stryer et al., 2004]. O raio de van der Waals é comumente utilizado em *softwares* de visualização gráfica de moléculas [Humphrey et al., 1996; Rego e Koes, 2015; Schrödinger, LLC, 2010] e também em cálculos de contatos [Kasahara et al., 2013; Kasahara e Kinoshita, 2014; Tanaka e Scheraga, 1975].

As pontes de hidrogênio ou ligações de hidrogênio são interações que possuem um componente eletrostático e que são formadas devido à diferença de eletronegatividade entre um átomo de hidrogênio (H) e um átomo mais eletronegativo – como flúor (F), oxigênio (O) e nitrogênio (N) – ao qual o hidrogênio se liga covalentemente

(doador de hidrogênio). Dada a diferença de eletronegatividade entre os dois átomos, a nuvem eletrônica do átomo de hidrogênio será deslocada para perto do átomo ligado covalentemente, criando, assim, um polo positivo no hidrogênio e um polo negativo no outro átomo. Nesse ponto, o hidrogênio tendo desenvolvido uma carga parcial positiva, poderá interagir de forma atrativa com um terceiro átomo também mais eletronegativo que o hidrogênio e com carga parcial negativa (aceptor de hidrogênio). Apesar de serem interações fracas, as pontes de hidrogênio são essenciais para muitas macromoléculas como as proteínas e os DNAs, além de serem responsáveis pelas propriedades peculiares e especiais da água [Kessel e Ben-Tal, 2010; Nelson e Cox, 2014; Stryer et al., 2004].

Empilhamentos aromáticos ou interações π-π, como o próprio nome diz são interações envolvendo moléculas que contenham grupos aromáticos (anéis aromáticos). Em proteínas, esse tipo de interação pode ocorrer a partir de 3 aminoácidos: Fenilalanina, Tirosina ou Triptofano. Anéis aromáticos possuem ligação dupla com ressonância (elétrons do orbital π são deslocalizados e distribuídos igualmente entre os átomos do anel), elétrons no orbital σ dentro do plano do anel e orbitais π acima e abaixo do plano (Figura 1.4). Os orbitais π fazem com que o anel adquira uma carga parcial negativa tanto no plano superior quanto no plano inferior e uma carga parcial positiva no plano do anel. Consequentemente, estas cargas parciais possibilitam que um anel interaja com outros anéis aromáticos ou átomos polares através de empilhamentos aromáticos [Kessel e Ben-Tal, 2010].

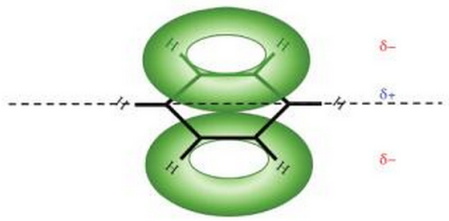


Figura 1.4. Distribuição de cargas em um anel aromático. No plano do anel (linha tracejada), a carga é parcialmente positiva. Acima e abaixo do plano, as cargas são parcialmente negativas devido ao deslocamento dos átomos do orbital π [Kessel e Ben-Tal, 2010].

O efeito de se adicionar um soluto apolar em um meio contendo água produz um tipo de interação entre moléculas apolares conhecido como efeito hidrofóbico ou interação hidrofóbica. Quando se analisa uma amostra contendo somente moléculas de água, é possível constatar diversas pontes de hidrogênio formadas entre estas moléculas – isto ocorre devido à grande diferença de eletronegatividade nos átomos de oxigênio e hidrogênio – formando, assim, uma estrutura interconectada bastante estável. Porém,

a adição de uma molécula qualquer (soluto) nesse meio causa uma perturbação na estabilidade dessa estrutura formada por moléculas de água. Quando se adiciona moléculas polares (soluto) na água, ocorre que as moléculas de água ao redor do soluto rompem suas pontes de hidrogênio para formar novas pontes de hidrogênio com o soluto. Dessa forma, as pontes de hidrogênio estabelecidas com o soluto compensam a perda de ligações locais entre as águas. Entretanto, o mesmo não ocorre entre uma molécula apolar e a água, uma vez que as interações entre o soluto e o solvente não irão acontecer. Ao invés disso, as moléculas de água se reorganizam e estabelecem pontes de hidrogênio entre si, produzindo, então, uma estrutura em forma de "jaula" ao redor do soluto. Adicionando-se mais moléculas apolares, ao invés de formar várias "jaulas" isoladas para cada molécula apolar, o que ocorre é a formação de uma única "jaula" que engloba todas as moléculas apolares. Assim, tendo os solutos apolares aversão às moléculas de água, eles tendem a se agregar e a interagir entre si como forma de evitar contatos com as moléculas de água [Dunn, 2010; Kessel e Ben-Tal, 2010; Levy e Onuchic, 2006; Metzler, 2003; Stryer et al., 2004].

Finalmente, as interações eletrostáticas são decorrentes de interações envolvendo átomos carregados. Sendo que átomos com cargas de mesmo sinal se repelem (interação repulsiva) e cargas com sinais diferentes se atraem (interação atrativa) tal como define a Lei de Coulomb. Pontes de hidrogênio também podem ser consideradas interações eletrostáticas, porém, envolvem átomos parcialmente carregados. Quando os átomos são totalmente carregados – ânions (átomos negativos) e cátions (átomos positivos) – essas interações são classificadas como iônicas, sendo que as interações envolvendo átomos de cargas diferentes são frequentemente referenciadas como pontes salinas [Branden e Tooze, 1999; Kessel e Ben-Tal, 2010; Nelson e Cox, 2014].

1.4 Ferramentas para cálculo e análise de interações proteína-ligante

Atualmente, existem algumas ferramentas que podem ser utilizadas para calcular e analisar interações proteína-ligante, porém, cada uma possui um foco diferente. Nesta seção, serão discutidas as ferramentas do estado da arte em termos de suas funcionalidades bem como prós e contras.

1.4.1 LIGPLOT

Em 1995, Wallace et al. [1995] publicaram o LIGPLOT que é uma ferramenta para gerar representações 2D das interações entre uma proteína e um ligante. Essa ferramenta possui a capacidade de transformar automaticamente o complexo proteína-ligante de uma representação tridimensional para uma representação em apenas 2 dimensões. Nesse processo, o LIGPLOT busca posicionar os átomos ou resíduos de forma que não haja sobreposição, prezando, portanto, pela legibilidade da imagem.

LIGPLOT é capaz de identificar ligações de hidrogênio e interações hidrofóbicas através do *software* HBPLUS [McDonald e Thornton, 1994]. Para calcular as ligações de hidrogênio, ele utiliza critérios de distância e angulação entre os átomos envolvidos. O critério define que a distância entre o hidrogênio e o átomo aceptor deve ser menor que 2,7Å; a distância entre o átomo doador e o átomo aceptor deve ser menor que 3,3Å; o ângulo entre esses três átomos deve ser maior que 90 °; e o ângulo entre o hidrogênio, o átomo aceptor e um átomo ligado ao átomo aceptor deve ser maior que 90 °. Já o critério utilizado para mapear interações hidrofóbicas define que a distância entre dois átomos de carbono deve ser menor que 3,9Å. Todos estes parâmetros podem ser personalizados pelo usuário.

Na representação em 2D, pontes de hidrogênio são representadas por linhas tracejadas entre dois átomos. Já os resíduos que estão envolvidos em interações hidrofóbicas são representados por um arco com pequenas linhas apontando em direção ao átomo do ligante com o qual ele estabelece a interação. O átomo do ligante, por sua vez, também possui um arco com linhas apontando em direção ao resíduo com o qual a interação hidrofóbica foi estabelecida.

Além dessas funcionalidades, LIGPLOT também permite que o usuário defina resíduos que não interajam diretamente com o ligante, mas que sejam importantes para a ligação. Como acontece no sítio ativo da enzima quimotripsina, em que o resíduo ASP102 não interage diretamente com o ligante, mas é responsável por orientar e estabilizar a HIS57 que atua ao lado da SER195 na hidrólise de ligações peptídicas do ligante [Nelson e Cox, 2014; Wallace et al., 1995]. Dessa forma, tais resíduos podem ser incluídos na representação 2D. Permite também que moléculas de água sejam adicionadas na representação. Finalmente, permite também a representação da acessibilidade de um átomo ao solvente através de um código de cores que pode ser personalizado.

1.4.1.1 Prós

LIGPLOT pode ser utilizado para qualquer complexo proteína-ligante de forma automática, permite adicionar informações de acessibilidade ao solvente, moléculas de água e resíduos que não interagem diretamente com o ligante.

1.4.1.2 Contras

Uma dificuldade encontrada no LIGPLOT é que não é possível realizar comparações entre os resultados desse *software*. Se o conjunto de dados consiste de uma proteína e vários ligantes, por exemplo, não é possível analisar as diferenças em cada complexo utilizando a própria ferramenta. Além disso, LIGPLOT não oferece análises estatísticas, visuais e interativas em larga escala que visam descrever os padrões, os tipos de interações e os mecanismos envolvidos no reconhecimento molecular de um conjunto de estruturas. Portanto, todas essas análises devem ser realizadas manualmente.

Ressalta-se também que nesse *software* estão disponíveis apenas as interações do tipo ponte de hidrogênio e interações hidrofóbicas.

1.4.2 LigPlot+

LigPlot+, como definido pelos próprios autores, é uma reformulação completa do LIGPLOT original [Laskowski e Swindells, 2011]. Nessa versão foram adicionadas interfaces gráficas para facilitar seu uso. A ideia básica desse *software* ainda é a mesma que o LIGPLOT, isto é, gerar uma representação 2D a partir da estrutura 3D do complexo. Os tipos de interação (pontes de hidrogênio e interações hidrofóbicas) e as funcionalidades disponíveis mantiveram-se as mesmas. A principal inovação dessa versão, é a possibilidade de comparar as representações 2D de diferentes complexos. Isto permitiu que estudos de diferentes conjuntos de dados sejam realizados, tais como aqueles que envolvem uma proteína e diversos ligantes ou várias proteínas e um ligante ou várias proteínas e vários ligantes.

Existem basicamente duas formas de comparação de resultados. Na primeira forma, todas as representações são posicionadas lado a lado. Já na segunda, todas as representações são sobrepostas com o objetivo de permitir a análise das principais diferenças entre cada representação. Em ambas as formas, resíduos que são equivalentes em diferentes estruturas são automaticamente destacados com um círculo. No LigPlot+, a equivalência entre resíduos é realizada através de um alinhamento de sequências utilizando o algoritmo de Needleman e Wunsch. Após esse alinhamento de sequências, LigPlot+ realiza a sobreposição das estruturas considerando o RMSD dos resíduos

equivalentes. Se o RMSD entre dois resíduos é maior que 3,5Å, eles são descartados e não são considerados equivalentes.

1.4.2.1 Prós

Como no LIGPLOT, LigPlot+ pode ser utilizado para qualquer complexo proteína-ligante de forma automática, permite adicionar informações de acessibilidade ao solvente, moléculas de água e resíduos que não interagem diretamente com o ligante.

Além disso, permite a comparação de diferentes complexos proteína-ligante e automaticamente seleciona resíduos equivalentes nas estruturas, que possibilita que o usuário identifique facilmente diferenças nas interações de cada complexo.

1.4.2.2 Contras

Apesar de poder analisar diferentes complexos e compará-los, essa funcionalidade ainda apresenta uma limitação com relação ao número de estruturas que se pode comparar simultaneamente. Por exemplo, tanto na análise por sobreposição quanto na análise por posicionamento lado a lado, analisar mais de 30 estruturas simultaneamente é uma tarefa bastante complexa. Além disso, LigPlot+ não oferece análises estatísticas, visuais e interativas em larga escala que visam descrever os padrões, os tipos de interações e os mecanismos envolvidos no reconhecimento molecular de um conjunto de estruturas. Portanto, todas essas análises devem ser realizadas manualmente.

Ressalta-se também que nesse *software* estão disponíveis apenas as interações do tipo ponte de hidrogênio e interações hidrofóbicas.

1.4.3 STING

STING [Mancini et al., 2004] é uma ferramenta *web* que dentre diversas outras funcionalidades, permite analisar os tipos de interações que dois átomos podem estabelecer através dos módulos *Graphical contacts* (GC) e *interface forming residue graphical contacts* (IFRgc). GC apresenta as interações entre átomos dos aminoácidos e IFRgc apresenta as interações entre uma proteína e todos os ligantes disponíveis na estrutura.

Os tipos de interações disponíveis nessa ferramenta são: pontes de hidrogênio, interações hidrofóbicas, interações carregadas (atrativas e repulsivas), empilhamentos aromáticos e pontes dissulfeto. A ferramenta detecta inclusive pontes de hidrogênio entre resíduos e ligantes que são intermediadas por moléculas de água. Essas interações são calculadas a partir de uma abordagem que considera os tipos de cada átomo

– definidas de acordo com as propriedades físico-químicas dos átomos – tal como proposto por Sobolev et al. [1999] e critérios de distância próprios. Por exemplo, por padrão, o STING define que interações hidrofóbicas somente ocorrem quando a distância necessária entre dois átomos está no intervalo entre 2 a 3.8Å e desde que os tipos destes átomos favoreçam essa interação, ou seja, que sejam átomos apolares. Os autores definem valores padrões para todas as interações, porém, o usuário pode redefini-las de acordo com sua preferência.

Tendo mapeado todas as interações, elas poderão, então, ser analisadas mais detalhadamente através de uma interface gráfica que lista as interações encontradas e seus tipos. Para realizar as análises o usuário deve definir um identificador PDB ou realizar o *upload* de uma estrutura, ou seja, no primeiro caso será utilizado o banco de dados PDB e no segundo caso o usuário pode escolher alguma estrutura armazenada em seu computador.

Uma restrição para o uso do STING é que o usuário deve ter o JAVA instalado em seu computador. Isto porque essa ferramenta utiliza o Jmol[1] (visualizador molecular para páginas *web*), que por sua vez requer que o JAVA esteja instalado no computador do usuário.

1.4.3.1 Prós

Permite a análise de interações entre átomos de dois aminoácidos ou interações proteína-ligante. STING pode ser utilizado para analisar qualquer estrutura, bastando que o usuário defina o identificador PDB ou faça o *upload* de arquivos PDB. STING também possui diversas outras funcionalidades, parâmetros personalizáveis e variados tipos de interações.

1.4.3.2 Contras

STING não permite que duas estruturas sejam comparadas diretamente e para isso o usuário deve analisar um resultado por vez. Apesar do STING apresentar resultados estatísticos, estes são voltados apenas para o estudo de uma única proteína. Sendo assim, os resultados estatísticos, visuais e interativos oferecidos por essa ferramenta não permitem a análise em larga escala, uma vez que o objetivo dessa ferramenta não é descrever os padrões, os tipos de interações e os mecanismos envolvidos no reconhecimento molecular de um conjunto de estruturas. Portanto, todas essas análises devem ser realizadas manualmente.

[1] http://jmol.sourceforge.net/

Além disso, outro ponto crítico se refere ao fato que o STING exige que seja instalado o *plugin* JAVA, conforme descrito no próprio portal do STING[2]. Consequentemente, os usuários não podem utilizar o navegador Chrome uma vez que o suporte a JAVA nesse navegador foi interrompido.

Por outro lado, mesmo tendo os navegadores que atendam essa exigência, os usuários ainda sofrem com um outro problema que se refere às políticas de segurança do JAVA. Devido à essas políticas, o JAVA deve ser configurado corretamente, caso contrário seu uso não será permitido pelo navegador [Rego e Koes, 2015].

1.4.4 CREDO

O CREDO é um banco de dados de interações proteína-ligante *web* [Schreyer e Blundell, 2009]. Sua grande vantagem em relação a muitos banco de dados de interações proteína-ligante, como os próprios autores avaliaram, é que essa ferramenta possui diversas funcionalidades para pesquisas avançadas e gratuitas. Por ser um banco de dados, todas as informações referentes às interações já foram pré-computadas e estão disponíveis para análises e pesquisas.

Nessa ferramenta, os átomos são classificados de acordo com a carga, conectividade e grupo funcional. Baseado no trabalho de Marcou e Rognan [2007], foram considerados os seguintes tipos atômicos: hidrofóbico, aromático, doador, aceptor, doador fraco, aceptor fraco, metal. Sendo que as interações possíveis a partir destes átomos são: interação hidrofóbica (dois átomos hidrofóbicos), empilhamento face a face (dois átomos aromáticos), empilhamento "ponta a face" (dois átomos aromáticos), ponte de hidrogênio (doador e aceptor ou aceptor e doador), ponte de hidrogênio fraca (qualquer combinação de átomos doadores fracos com aceptores fracos), interação π-cátion (cátion e aromático ou aromático e cátion), complexo metálico (metal e aceptor). Nessa lista de interações, o primeiro átomo de cada combinação sempre pertence a um aminoácido e o segundo a um ligante. Por exemplo, no complexo metálico o átomo classificado como metal pertence à proteína. Já as interações iônicas foram definidas utilizando os tipos de átomo Ionizável positivamente e Ionizável negativamente extraídos do RDKit[3]. Outras interações também identificadas são: interações de van der Waals, colisões de van der Waals (van der Waals *clashes*), interações covalentes e ponte de halogênio. Essa última interação é definida segundo os átomos doadores de halogênio e aceptores de halogênio. Essas classificações foram utilizadas inclusive para todos os estados de protonação dos átomos dos aminoácidos, utilizando as informações contidas no *site*

[2]http://www.cbi.cnptia.embrapa.br/SMS/STINGm/help/GS_requirements_jmol.html
[3]http://www.rdkit.org/

do PDB. Todos os tipos de átomos e interações entre ligantes e proteínas foram pré-calculadas considerando os tipos de cada átomo, bem como critérios de distância. A representação das interações foi feita através de um *fingerprint* binário chamado SIFt [Brewerton, 2008; Deng et al., 2004] em que cada bit indica se há ou não uma interação. Esse método é útil, pois, possibilita a comparação de similaridade das interações estabelecidas em vários complexos. Além das interações estabelecidas de forma direta entre átomos da proteína e ligante, há também a detecção de pontes de hidrogênio entre resíduos e ligantes que são intermediadas por moléculas de água.

Outra informação disponível no CREDO é o mapeamento de todas as cadeias em suas sequências do UniProt [Bateman et al., 2015]. Segundo os autores, essas informações permitem que sejam avaliados os efeitos ocasionados na estrutura da proteína e sua função quando há variações na sequência.

CREDO também utiliza uma abordagem para identificação de ligantes promíscuos utilizando o banco de dados para classificação de proteínas SCOP [Murzin et al., 1995]. Nessa abordagem, os ligantes são considerados promíscuos quando eles são encontrados em contato com pelo menos duas famílias de proteínas diferentes no SCOP.

CREDO permite a pesquisa de ligantes similares através de uma abordagem utilizando *fingerprints* (Seção 2.9.1). A similaridade entre dois *fingerprints* F1 e F2 é calculada segundo o coeficiente de Tanimoto:

$$T(F1, F2) = \frac{c}{a + b + c} \tag{1.1}$$

Onde c é o número de posições idênticas e iguais a 1 em ambos os *fingerprints* de F1 e F2, a é o número de posições em que os bits são iguais a 1 somente no *fingerprint* de F1 e b é o número de posições em que os bits são iguais a 1 somente no *fingerprint* de F2. Além desse tipo de pesquisa, o usuário ainda tem a possibilidade de comparar ligantes e suas conformações através de similaridade da forma, permitindo, assim, que sejam analisados o relacionamento entre conformações ativas e aquelas não ativas. Para isso, eles utilizam o algoritmo *Ultrafast Shape Recognition* (USR) proposto por Ballester e Richards [2007]. Essas funcionalidades são úteis, por exemplo, para técnicas de triagem virtual em que frequentemente é necessário pesquisar compostos que sejam similares a um ligante conhecido. Métodos de triagem virtual consistem na identificação de novos compostos que possam desempenhar uma função terapêutica frente a uma molécula alvo por meio de métodos computacionais e a partir de uma grande quantidade de ligantes [Lavecchia e Giovanni, 2013; Verli, 2014].

Outra funcionalidade disponível é a possibilidade de fazer pesquisas através de requisições diretamente ao *web service*. Assim, o usuário pode usufruir dos serviços

disponíveis sem a necessidade de utilizar a página *web*. O usuário pode, por exemplo, implementar um algoritmo em seu computador que utilize as pesquisas de similaridade do CREDO através de requisições a esse servidor.

1.4.4.1 Prós

O CREDO como banco de dados, apresenta diversas informações sobre propriedades químicas dos ligantes, proteínas e interações proteína-ligante. A ferramenta apresenta diversas funcionalidades que auxiliam na pesquisa e descoberta de moléculas similares a um determinado ligante. CREDO ainda possui informações de um amplo escopo de tipos de átomos e interações mapeadas em seu banco de dados e ainda inclui informações sobre interações proteína-ligante intermediada por moléculas de água. CREDO ainda relaciona as informações das proteínas com outros banco de dados como UniProt [Bateman et al., 2015], SCOP [Murzin et al., 1995] e Pfam [Finn et al., 2014]. Outra função útil é a possibilidade de realizar requisições diretamente ao *web service*.

1.4.4.2 Contras

CREDO não oferece análises estatísticas, visuais e interativas em larga escala que visam descrever os padrões, os tipos de interações e os mecanismos envolvidos no reconhecimento molecular de um conjunto de estruturas. Portanto, todas essas análises devem ser realizadas manualmente.

Outro problema é que a base de dados depende de atualizações das informações da base à medida que novos complexos vão sendo disponibilizados no site do PDB.

1.4.5 GIANT

Ao contrário das outras ferramentas já descritas que focam em determinar os tipos de interações proteína-ligante, O GIANT tem seu foco voltado para a descrição dos modos de ligação dos ligantes com a proteína [Kasahara e Kinoshita, 2014]. Ele é um servidor *web* para analisar interações proteína-ligante com base nos padrões atômicos obtidos por análises estatísticas dos complexos. Através destes padrões é possível constatar as preferências espaciais das interações entre os átomos da proteína e do ligante.

O servidor é composto de uma base de dados pré-calculada que foi obtida a partir de um trabalho anterior dos mesmos pesquisadores [Kasahara et al., 2013]. Nesse trabalho, foram utilizadas 23.040 estruturas do PDB. Os resíduos de todas as proteínas foram decompostas em fragmentos de três átomos ligados covalentemente e os tipos de

cada átomo dos fragmentos foram definidos de acordo com o elemento e sua posição no resíduo. Por exemplo, na cadeia principal tem-se os átomos classificados como carbono (C) e carbono-α (Cα) e na cadeia lateral os átomos carbono-β (Cβ) e carbono-γ (Cγ). Os átomos dos ligantes foram classificados de acordo com a tabela do *Tripos force field* [Clark et al., 1989]. Essa tabela classifica os átomos de acordo com o número de interações estabelecidas, hibridização, aromaticidade e outras propriedades estruturais.

Em seguida, foram coletadas todas as interações entre fragmentos dos resíduos e átomos do ligante. Os contatos nessa abordagem foram definidos utilizando um critério de distância que estabelece que dois átomos estão em contato se a distância entre eles for menor que a soma de seus raios de van de Waals mais um limiar de distância igual a 1Å. A partir das coordenadas dos fragmentos (coordenadas dos três átomos) foram geradas as distribuições espaciais das interações em um plano tridimensional.

Em seguida, os autores aplicaram uma abordagem estatística para o reconhecimento de padrões de acordo com a distribuição espacial realizada. Os autores aplicaram o modelo de misturas Gaussianas (*Gaussian mixture model* ou GMM) que é um modelo probabilístico utilizado para detectar classes em um conjunto de dados de forma não supervisionada, isto é, sem conhecimento prévio sobre as classes existentes nesse conjunto. Nesse contexto, os padrões podem ser considerados como uma classe [Bishop, 2006]. O modelo utilizado por Kasahara et al. [2013] define a função de densidade probabilística como sendo:

$$p(x) = \sum_{k=1}^{K} \pi_{\mathrm{k}} N(x|\mu_{\mathrm{k}}, \Sigma_{\mathrm{k}}) \tag{1.2}$$

Onde p(x) descreve a probabilidade de que um átomo no ponto x interaja com um fragmento da proteína em uma dada região, $N(x|\mu_{\mathrm{k}}, \Sigma_{\mathrm{k}})$ indica uma função Gaussiana com média μ_{k}, matriz de covariância Σ_{k} e π_{k} é o peso da função Gaussiana no GMM. Os parâmetros μ_{k}, Σ_{k} e π_{k} devem ser estimados de forma estatística para maximizar a probabilidade p(x), isto é, para que a função p(x) tenha a capacidade de definir se um ponto x pertence ou não a uma dada classe com uma alta probabilidade. Para isso, Kasahara et al. [2013] utilizaram a abordagem variacional Bayesiana proposta por Attias [1999]. Com essa abordagem de reconhecimento, foram reconhecidos 13.519 padrões de interação.

Após reconhecer os padrões, Kasahara et al. [2013] associaram as interações a cada padrão. Isto porque no processo anterior foram identificados apenas os padrões, ou seja, as interações ainda não haviam sido classificadas em nenhum padrão. Para tanto, calculou-se a distância de Mahalanobis entre um ponto x e uma componente Gaussiana g (padrão). Sendo que uma interação somente é associada a uma com-

ponente Gaussiana quando essa distância é menor que 2,5. A distância é dada pela seguinte fórmula:

$$D(x, g) = \sqrt{(x - \mu_g)^T \Sigma_g{}^{-1} (x - \mu_g)} \tag{1.3}$$

Utilizando essa abordagem há a possibilidade de que uma interação seja associada a mais de um padrão. Assim, 63,4% das interações foram associadas em pelo menos 1 padrão.

Com base nessas análises, Kasahara e Kinoshita [2014] construíram uma ferramenta *web* chamada GIANT. Essa ferramenta tem as seguintes funcionalidades: *Analisador de complexo* (*Complex analyzer*) que permite que o usuário inspecione as interações proteína-ligante com base nos padrões de interação obtidos por Kasahara et al. [2013], e *Visualizador de padrões* (*Pattern viewer*) que mostra um resumo de cada padrão de interação.

No Analisador de complexo, o usuário poderá ver quais são os resíduos e átomos do ligante que estabelecem interações, e quais são os pares de fragmentos e átomos do ligante que estão interagindo. O usuário poderá ainda ver o sítio de ligação e as interações estabelecidas através de uma visualização similar ao PyMOL [Schrödinger, LLC, 2010]. Já no Visualizador de padrões o usuário poderá analisar a distribuição de probabilidade de cada padrão de interação ao longo do espaço 3D. Para isso, um fragmento de aminoácido é centralizado e em torno desse fragmento são mostradas diversas malhas que correspondem aos padrões de interação. Regiões dentro destas malhas são posições estatisticamente preferíveis dos átomos do ligante para interagir com o fragmento. A partir dessa funcionalidade, os usuários poderão ainda verificar quais ligantes e quais complexos proteína-ligante também possuem um determinado padrão.

O GIANT permite também que o usuário utilize tanto os complexos já definidos no banco de dados quanto um complexo local em seu computador, para isso o usuário deve fazer o *upload* do arquivo PDB desejado.

Entretanto, como o GIANT utiliza o Jmol (visualizador molecular para páginas *web*) o usuário deve ter instalado o JAVA em seu computador.

1.4.5.1 Prós

GIANT aplica uma abordagem diferente para a análise de complexos proteína-ligante, focando nos modos de ligação dos ligantes com as proteínas. Dessa forma, é possível constatar as preferências espaciais das interações entre os átomos da proteína

e do ligante. Possibilita ainda que sejam encontrados outros complexos que tenham o mesmo padrão.

Além disso, permite que o usuário análise complexos que estão no seu computador através do *upload* de arquivos.

1.4.5.2 Contras

GIANT não define quais são os tipos de interações especificamente, apenas define que há contatos entre os átomos. Por exemplo, não é possível saber se uma interação é ponte de hidrogênio ou se é um empilhamento aromático.

Além disso, essa ferramenta também não oferece análises estatísticas, visuais e interativas em larga escala que visam descrever os padrões, os tipos de interações e os mecanismos envolvidos no reconhecimento molecular de um conjunto de estruturas. Portanto, todas essas análises devem ser realizadas manualmente.

Como ocorre com a ferramenta STING, o GIANT também requer que o usuário tenha o JAVA devidamente instalado no computador. Isto porque a ferramenta utiliza o visualizador molecular Jmol que depende da plataforma JAVA.

1.4.6 Sumário de prós e contras

Na Seção 3.1, iremos apresentar um sumário de prós e contras de cada ferramenta aqui apresentada, bem como os tipos de interações que cada uma possui disponível. Preferimos incluir essas informações apenas na seção de Resultados, pois primeiramente iremos propor e apresentar as funcionalidades e metodologias de nossa ferramenta. Assim, será possível realizar um comparativo entre nossa ferramenta e as demais.

1.5 Visualização de dados

Segundo Schroeder et al. [2001], com o surgimento de novos métodos e tecnologias constantemente sendo propostos para responder às mais diversas perguntas nas áreas da biologia e química, uma quantidade de dados enorme decorrente dessa evolução científica vem sendo produzida. Experimentos tais como a expressão de genes em microarranjos, por exemplo, produzem um complexo e volumoso conjunto de informações, logo, a análise manual destes dados por um biólogo se torna proibitiva.

A computação se mostrou uma grande aliada nas pesquisas científicas, pois permite a produção, recuperação, análise e interpretação de grandes conjuntos de informações em uma quantidade de tempo inferior àquela que um ser humano levaria.

Diversas bases de dados que visam armazenar e tornar públicas as novas descobertas científicas foram propostas nos últimos anos, sendo que algumas delas dispõem informações sequenciais, estruturais de proteínas e interações entre proteínas e outras moléculas (Uniprot [Bateman et al., 2015], PDB [Berman et al., 2000], Pfam [Finn et al., 2014] , Relibase [Hendlich et al., 2003], PROCOGNATE [Bashton et al., 2008]).

Não obstante, uma dificuldade que surge a partir desse rápido crescimento de informações disponíveis está relacionado a como interpretá-las e como utilizá-las da melhor forma possível, a fim de responder diversas questões propostas.

De fato, segundo Few [2009] o crescimento rápido da quantidade de informação disponível não é um real problema. A maior dificuldade está em como obter, explorar e analisar essa imensa quantidade de dados disponíveis e extrair valiosas informações, conhecimentos ou mesmo novas indagações. Ainda, de acordo com Schroeder et al. [2001], esse crescimento exponencial de informações resulta em uma constante necessidade de produção de melhores métodos para analisar e representar as informações disponíveis. Assim, um desafio enfrentado se refere à escolha dos métodos e técnicas que serão empregadas para transformar as informações em uma forma visual que seja compreensível.

Segundo Few [2009], existem diversas formas de representar as informações, porém, alguns tipos de representações combinam naturalmente com determinados tipos de dados. Ainda, segundo Saraiya et al. [2005] visualizações bem planejadas e desenvolvidas com o intuito de possibilitar a máxima compreensão das informações podem ser tão eficientes quanto sistemas mais complexos.

Nesse quesito, a visualização de dados tem se tornado uma área de estudo bem estabelecida nos últimos anos e é crescente seu uso na representação de dados biológicos. Isto se deve ao fato de que tais aplicações se mostram capazes de auxiliar as análises e interpretações de um grande conjunto de informações, provendo a base para explorar e explicar os dados de forma interativa e eficiente [Saraiya et al., 2005; Wong, 2012]. Além disso, a visualização de dados favorece e encoraja o uso de reconhecimento de padrões tais como dispersão dos dados, simetria, agrupamentos e lacunas.

1.5.1 O que é visualização de dados?

Em termos gerais, segundo Few [2009], visualização de dados envolve todos os tipos de representações visuais que auxiliam na exploração, análise e comunicação dos dados e informações.

Representações visuais permitem que padrões, tendências e exceções sejam facilmente identificáveis e se tornem inteligíveis. Além disso, o grande potencial das repre-

sentações visuais se dá pelo fato de que elas ampliam nossa capacidade de memorização, de forma que informações como números ou textos que são por vezes complexos e difíceis de serem analisados isoladamente, se tornem facilmente reconhecíveis e estejam disponíveis na memória. Segundo Few [2009], representações de forma visual de informações quantitativas, por exemplo, aumentam consideravelmente nossa capacidade de pensar e interpretar as mesmas.

1.5.2 Ferramentas e técnicas para visualização de dados

Em técnicas de visualização de dados, frequentemente, incluem-se os bem conhecidos gráficos de barras, pizza, linhas, pontos, entre muitos outros tipos de visualizações. Grafos também são bastante utilizados em visualizações, por exemplo, para representar redes de interações de proteínas em que cada nó representa uma proteína e cada aresta representa uma interação, como pode ser visto pela figura 1.5.

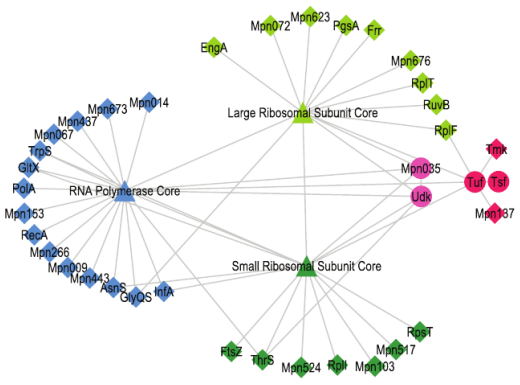


Figura 1.5. Exemplo de uso de grafos para visualizações de dados biológicos [Gehlenborg et al., 2010].

Dentre tantas opções de visualizações, pode-se dizer que a escolha entre uma e outra deve refletir na clareza e eficácia das informações representadas, propiciando, assim, o entendimento do usuário em relação a essas informações e também contribuir na formulação de novas questões. Entretanto, uma maneira de propiciar uma maior efetividade nas análises é através do uso de interatividade com o usuário.

Nesse quesito, Few [2009] aborda diversas técnicas que podem ser aplicadas a fim de permitir essa maior interatividade nas análises de dados. Dentre elas, pode-se citar técnicas que permitam a comparação de informações; opções para ordenação de valores seja de forma alfabética ou numérica e de forma crescente ou decrescente; opções que permitam ampliar e aproximar a visão para uma parte em específico da visualização (zoom); filtragem de informações com o propósito de reduzir o número de informações

sendo visualizadas a um conjunto menor para facilitar a análise do que se pretende estudar e tirar o foco de elementos não importantes em um certo momento; detalhes sob demanda, que consiste em oferecer opções para que o usuário obtenha informações detalhadas sobre cada elemento da visualização sempre que necessário.

Um exemplo prático de uma ferramenta web recentemente proposta e que aplica tais técnicas de visualização de dados para auxiliar a busca por novos conhecimentos, é a ferramenta que propusemos em Silveira et al. [2014]. Nesse trabalho, aplicamos diferentes conjuntos de informações sobre resíduos de aminoácido e mutações na estrutura primária da proteína (sequência de aminoácidos) com o propósito de permitir a identificação de mutações importantes e suas possíveis consequências para a função de uma proteína. A figura 1.6, ilustra um exemplo de visualização na qual o usuário tem a possibilidade de verificar a conservação de resíduos em uma família de proteínas a partir de resultados de um alinhamento de sequências baseado em estrutura e ainda permite obter detalhes sobre características físico-químicas e topológicas de forma interativa e sob demanda.

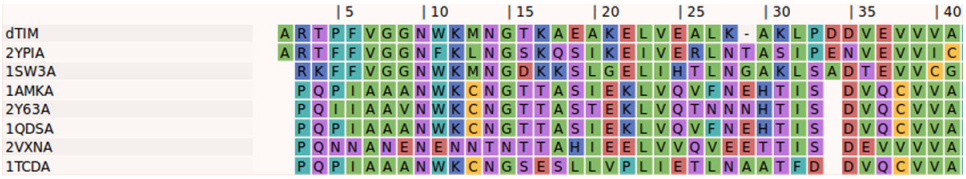


Figura 1.6. Exemplo de uma ferramenta web para a análise de mutações em proteínas [Silveira et al., 2014].

Portanto, técnicas de visualização se mostram uma poderosa ferramenta para explorar padrões, tendências e/ou correlações entre os dados, uma vez que visualizações bem desenvolvidas descrevem as informações e permitem inferências de forma simples, substituindo uma árdua e proibitiva análise manual de uma grande extensão de informações textuais ou numéricas [Wong, 2012]. Além disso, através de técnicas simples, pode-se aumentar consideravelmente a maneira com a qual o usuário interage com as informações, propiciando encontrar o conhecimento e as respostas que se busca de forma concisa, eficaz, interativa e até mesmo divertida.

1.6 Motivação

Elucidar os mecanismos envolvidos no reconhecimento molecular e quais forças contribuem para o reconhecimento é um problema central na biologia [Kasahara e Kinoshita, 2014], uma vez as interações entre duas moléculas são de extrema importância

para os sistemas biológicos como um todo. As diversas funções desempenhadas por proteínas são, em grande parte, dependentes de sítios específicos de ligação e interação com outras moléculas. Essas moléculas, então, se ligam à proteína receptora através de uma complementariedade molecular que geralmente ocorre de forma a obter uma conformação que seja energeticamente favorável [Böhm, 2005]. Portanto, o estudo das interações proteína-ligante permite um maior entendimento sobre como ocorre a seletividade e quais são as forças essenciais para que o reconhecimento ocorra.

Atualmente, existem diversas ferramentas para auxiliar a análise de interações proteína-ligante através de diferentes abordagens. Entretanto, nenhuma dessas ferramentas apresenta análises estatísticas, visuais e interativas em larga escala para o estudo de interações proteína-ligante. Em geral, essas ferramentas não oferecem mecanismos para comparações de complexos através de dados estatísticos, tais como a frequência de resíduos que mais interagem com os ligantes, tipos de átomos ou interações. Dessa forma, o usuário deve realizar suas análises de forma manual.

É interessante oferecer mecanismos que possibilitem ao usuário comparar diversos conjuntos de dados constituídos seja por uma proteína e vários ligantes ou várias proteínas e um ligante. Entender as diferenças e semelhanças em tais conjuntos é essencial para elucidar os mecanismos envolvidos no reconhecimento molecular nesses conjuntos de dados. Esse estudo, por exemplo, contribui para as pesquisas farmacológicas, pois compreender esses mecanismos auxilia na formulação de novos compostos químicos para uso terapêutico.

Portanto, propomos uma ferramenta que visa ser fácil e intuitiva a partir do uso de estratégias visuais para descrever os padrões e tipos de interações estabelecidas entre proteínas e seus ligantes. Nosso foco é proporcionar meios para que o usuário realize suas análises de interações proteína-ligante de forma estatística, visual e interativa em larga escala. Possibilitando, por conseguinte, comparar semelhanças e diferenças nas interações estabelecidas em diferentes conjunto de dados. Para tanto, propomos o nAPOLI (Analysis of PrOtein Ligand Interactions), uma ferramenta *web* interativa para o estudo em larga escala das interações proteína-ligante utilizando estratégias visuais e análises estatísticas. A ferramenta pode ser acessada em: http://www.napoli.dcc.ufmg.br/

1.7 Objetivos

1.7.1 Objetivo geral

Desenvolver uma ferramenta *web* para permitir a análise de padrões de interação proteína-ligante para quaisquer conjunto de estruturas PDB seja quando há uma mesma proteína e diversos ligantes diferentes, quando há várias proteínas diferentes e um único ligante ou quando há várias proteínas e vários ligantes diferentes. Essa ferramenta deve ser capaz de permitir análises e comparações de forma fácil e intuitiva através de dados estatísticos e técnicas de visualização de dados que proporcionem interatividade e melhorem a experiência do usuário com a ferramenta.

1.7.2 Objetivos específicos

- Fazer um levantamento e estudo sobre o estado da arte;
- Investigar quais tipos de interações modelar;
- Definir as filtragens que serão feitas nos arquivos PDBs;
- Estudar o critério de distância adotado no cálculo de contato de acordo com a literatura;
- Calcular contatos de acordo com o critério geométrico;
- Projetar e implementar o cálculo de interações de acordo com os tipos atômicos e critérios de distância;
- Modelar e implementar as interações proteína-ligante como grafos;
- Projetar e implementar os algoritmos necessários para gerar os resultados das análises de forma automática de acordo com requisições dos usuários;
- Projetar e implementar algoritmos que busquem e encontrem regiões contendo grupos de ligantes nas proteínas;
- Investigar e implementar técnicas de visualização de dados adequados aos dados e ao problema;
- Projetar e implementar todas as páginas *web*;
- Implementar o servidor *web* e o banco de dados;
- Realizar um estudo de caso.

Capítulo 2

Materiais e métodos

2.1 Projeto e implementação do sistema nAPOLI

Nesta seção, serão descritas as decisões de projeto e as linguagens utilizadas para implementação do sistema.

O servidor foi construído utilizando a linguagem PHP[1] e CodeIgniter[2] que é um *framework* para PHP que auxilia e simplifica a construção de páginas PHP. Um *framework* em geral se refere a um conjunto de algoritmos que oferecem funcionalidades genéricas para agilizar e simplificar a produção de algoritmos. O desenvolvimento das páginas também empregou as linguagens HTML5 para definir a estrutura e a apresentação das páginas web, *Cascading Style Sheets* (CSS) para estilizar páginas web e JavaScript para proporcionar conteúdo dinâmico às páginas. Nesse contexto, foram utilizadas as seguintes bibliotecas JavaScript: jQuery[3] é utilizada para simplificar a criação de conteúdo dinâmico; D3.js[4] [Bostock et al., 2011] é utilizada como recurso para a criação de gráficos, tabelas e aplicações baseadas em técnicas de visualização de dados; DataTables[5] é utilizada para a criação de tabelas dinâmicas; 3Dmol.js[6] [Rego e Koes, 2015] é utilizada para a criação de visualizações tridimensionais das estruturas de proteínas com várias funcionalidades interativas. Já o *framework* Bootstrap[7] foi utilizado tanto para a implementação do conteúdo HTML, quanto na estilização do site (CSS) e na produção de conteúdo dinâmico (JavaScript).

[1] http://www.php.net
[2] http://www.codeigniter.com/
[3] https://jquery.com/
[4] http://d3js.org/
[5] https://datatables.net/
[6] http://3dmol.csb.pitt.edu/index.html
[7] http://www.getbootstrap.com

A base de dados utilizada para armazenar todas as informações da aplicação foi o MySQL.

Quanto ao sistema local, foram utilizados as linguagens PERL, Java e Python. A linguagem PERL é a linguagem principal do sistema, tendo sido escolhida dado sua simplificidade e rapidez para a implementação de conteúdo que faz uso intenso de processamento de textos como, por exemplo, leitura de arquivos PDB. Python e Java tiveram seu uso mais restrito às ferramentas de terceiros que requisitaram o uso destas linguagens como, por exemplo, o *software* PyMOL [Schrödinger, LLC, 2010] foi utilizado para gerar as imagens da sobreposição de grupos de ligantes (Figura 2.12(a)) de forma automática e para isso foi necessário implementar um algoritmo em Python posto que o PyMOL apenas suporta scripts escritos nessa linguagem.

2.2 Arquivos PDBs

O Protein Data Bank (PDB) [Berman et al., 2000] é o maior banco de dados contendo estruturas de proteínas na atualidade [Wang et al., 2011]. Esse banco apresenta hoje (acessado em 05/06/2015), 109.274 arquivos com estruturas de macromoléculas depositadas que incluem além de estruturas proteicas, informações e estruturas de ácidos nucleicos e outros complexos. Estruturas de complexos proteína-ligante encontradas no PDB são informações valiosas, pois, apresentam em nível atômico as interações moleculares entre proteínas e ligantes. Isso, portanto, permite a investigação dos processos e forças envolvidas no reconhecimento molecular entre duas moléculas.

Um arquivo de estrutura ou arquivos PDB são representados em um formato específico e padronizado, chamado formato PDB[8]. Um arquivo PDB pode ser dividido em duas partes principais, a primeira contêm informações e detalhes sobre a estrutura e a segunda contêm as coordenadas dos átomos. Na primeira seção, informações tais como o método utilizado para resolver a estrutura da proteína, o nome dos autores que resolveram a estrutura, nome da macromolécula e a lista de cadeias disponíveis na estrutura, informações sobre as estruturas secundárias, entre tantas outras informações estão disponíveis. A segunda seção, que para nós é considerada a mais importante, contêm informações relativas às coordenadas dos átomos da proteína e dos ligantes (registros chamados de ATOM e HETATM), e informações sobre ligações covalentes e conexões entre átomos (registros chamados de CONECT). A partir destas coordenadas, é possível calcular os tipos atômicos e interações envolvidas entre proteínas e seus ligantes.

[8]http://www.wwpdb.org/documentation/file-format-content/format33/v3.3.html

2.3 Visão geral da ferramenta

O nAPOLI pode ser dividido basicamente em duas etapas distintas: a pesquisa de grupos de ligantes e a produção dos resultados para a análise de padrões de interação proteína-ligante. A primeira etapa consiste na composição de uma base de dados constituída seja por uma proteína e vários ligantes, várias proteínas e um ligante ou várias proteínas e vários ligantes. Já a segunda etapa, consiste na produção dos resultados para a base de dados definida pelo usuário.

A Figura 2.1(a) a seguir demonstra os passos utilizados para encontrar grupos de ligantes de acordo com as regiões onde estes se encontram. Estes passos serão descritos na Seção 2.4. A lista de PDBs nessa etapa são apenas os identificadores dos arquivos PDB encontrados na primeira etapa da pesquisa de ligantes (ver Seção 2.4).

A Figura 2.1(b) a seguir, demonstra os principais passos utilizados para produzir os resultados das análises de padrões Proteína-ligante. Cada etapa será descrita mais detalhadamente ao longo deste capítulo. Todos os parâmetros definidos pelo usuário são também utilizados ao longo da execução do nAPOLI. Ao final, todos os resultados gerados estarão disponíveis ao usuário através do site do nAPOLI.

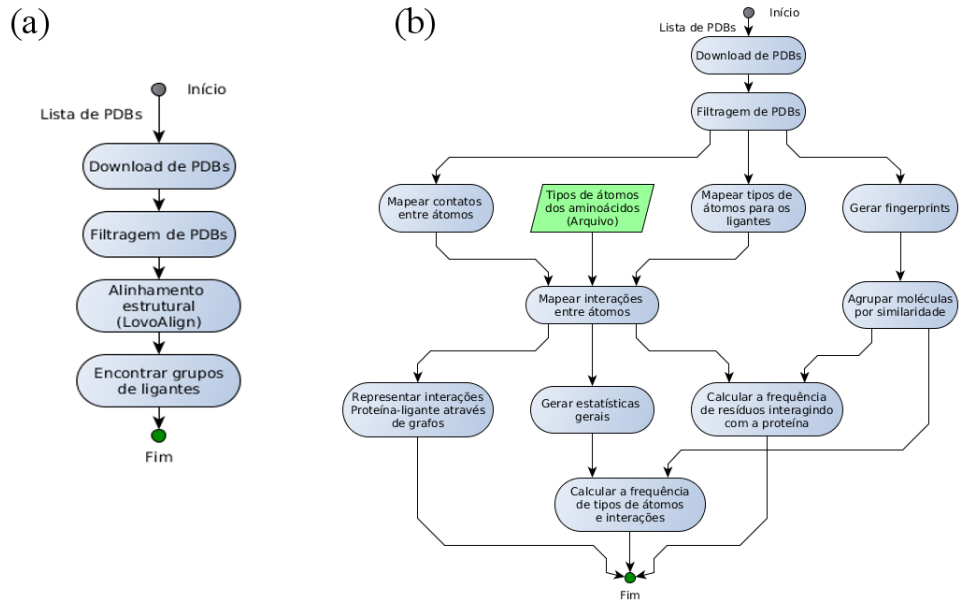


Figura 2.1. Diagramas de fluxo do sistema. (a) Visão geral das etapas para encontrar grupos de ligantes de acordo com as regiões da proteína que estes estejam localizados. (b) Visão geral das etapas para produzir os resultados das análises de padrões Proteína-ligante.

2.4 Primeiros passos: a pesquisa de ligantes

O primeiro passo para qualquer pesquisador, antes de iniciar seus experimentos, é encontrar uma base de dados que satisfaça aos objetivos de seu estudo. As etapas de pesquisa exigem um planejamento prévio, um conhecimento sobre quais informações se pretende utilizar, a coleta cuidadosa de dados, possivelmente um pré-processamento para limpeza e padronização e finalmente seu uso em si. Nestas primeiras etapas, é importante uma pesquisa de caráter exploratório em que o pesquisador possa identificar novas bases de dados, filtrar dados que realmente sejam de seu interesse e finalmente encontrar novas informações que permitam obter as respostas de que precisa.

Tendo isso em vista, o nAPOLI disponibiliza uma ferramenta de pesquisa simples que oferece diversas opções ao pesquisador para encontrar novas bases de dados que sejam de seu interesse. Por ser uma ferramenta que tem como objetivo ajudar na busca por padrões de interações entre proteína e ligante, essa etapa fornecerá ao pesquisador uma forma simples de encontrar complexos de proteínas e ligantes para a sua pesquisa. Muitas vezes, o pesquisador possui uma proteína de interesse relacionada à sua pesquisa e deseja encontrar novas proteínas que tenham uma certa similaridade à proteína de interesse, além de possuírem um ou mais ligantes interagindo com suas cadeias de aminoácidos. Dessa forma, a ferramenta busca solucionar esse tipo de demanda e auxiliar a quem deseja encontrar novos alvos para a sua pesquisa ou mesmo apenas adquirir mais informações sobre a conservação de resíduos/interações nos complexos.

Resumidamente, nessa primeira etapa o objetivo será encontrar PDBs que possuam uma certa similaridade de sequência com uma proteína alvo e que disponham de ligantes interagindo com a proteína. Em seguida, serão descritas todas as etapas necessárias para realizar essa pesquisa na base de dados PDB.

2.4.1 Pesquisando PDBs

Nesta primeira etapa, a pesquisa se inicia com o usuário definindo um identificador PDB e uma cadeia válida para esse PDB, além de definir um valor de identidade de sequência que será utilizado para calcular a similaridade entre as proteínas. Essas três informações são essenciais para o sucesso da pesquisa. Por exemplo, se alguém informar um identificador de um PDB (PDB *id*) ou cadeia (*chain*) que seja inválida, não será possível encontrar nenhum outro PDB e o usuário deverá iniciar novamente essa etapa da pesquisa passando um novo identificador.

Para realizar essa pesquisa, o sistema utiliza o *web service* do PDB [9]. Nesse servi-

[9] http://www.rcsb.org/pdb/software/rest.do

dor, estão disponíveis diversas opções de pesquisa avançada para recuperação de estruturas de proteínas, ácidos nucleicos e outros complexos. Uma destas funções permite aos usuários pesquisarem proteínas e ácidos nucleicos por similaridade de sequências utilizando os algoritmos BLAST [Altschul et al., 1990], FASTA [Pearson e Lipman, 1988] ou PSI-BLAST [Altschul et al., 1997]. Função esta que pode ser escolhida na própria página do site PDB [10] na seção de *Pesquisas avançadas (Advanced Search)*, escolhendo o tipo de pesquisa nomeada como *Sequence (BLAST/FASTA/PSI-BLAST)*. Dessa forma, para dispor das mesmas funções disponíveis desse servidor, foram adicionadas as seguintes opções:

- PDB *Id*: o identificador correspondente à estrutura proteica a ser utilizada na pesquisa.
- *Chain*: a cadeia de aminoácidos presente na estrutura definida pelo PDB *id*.
- *Sequence identity cutoff* (%): filtro para remover entradas que possuam baixa similaridade de sequências.
- *Search tool*: o algoritmo a ser utilizado no cálculo de similaridade de sequências. O padrão para esse campo é o algoritmo BLAST conforme o próprio site do PDB.
- *Mask low complexity*: filtro para mascarar regiões de baixa complexidade em uma sequência com o objetivo de filtrar alinhamentos hipotéticos. Estas regiões na sequência de consulta equivalem a valores iguais ao caractere X. O padrão para esse campo é *Yes* conforme o próprio site do PDB.
- *E-value cutoff*: *E-value* ou *Expected value* é um parâmetro que corresponde ao número de distintos alinhamentos com uma pontuação igual ou melhor e que são esperados de ocorrer por acaso, ou seja, aleatoriamente, na busca por sequências similares [Verli, 2014]. O valor padrão para esse campo é 10 conforme o próprio site do PDB.

A única opção que não disponibilizamos nessa versão é a busca por sequências. Quando a pesquisa é feita por sequência, não é configurado nenhuma estrutura para ser utilizada como modelo para os alinhamentos. Por padrão, quando isso ocorre, nosso algoritmo utiliza o primeiro PDB, em ordem alfabética, como modelo. Assim, essa estrutura escolhida pode não representar bem as demais estruturas na base de dados do usuário. Isso interferiria na busca por grupos de ligantes e na contagem de resíduos

[10] http://www.rcsb.org/pdb/search/advSearch.do?search=new

que mais frequentemente interagem com os ligantes. Como consequência, os resultados retornados ao usuário poderiam não ser relevantes. Esse problema pode ocorrer principalmente quando o usuário utiliza valores de identidade muito baixos. Com 100% de identidade, por exemplo, a escolha por ordem alfabética não faria diferença uma vez que todas estruturas teriam sequências iguais. Então, para evitar essas disparidades de resultado preferimos remover temporariamente essa opção da ferramenta. Nas próximas versões, essa função será incluída e estará disponível para uso.

Após definir todos os parâmetros desejados, considerando que três deles são obrigatórios (PDB *id*, *Chain* e *Sequence identity cutoff*), o usuário irá submeter a pesquisa e a ferramenta buscará por um conjunto de identificadores PDB que satisfaçam os valores definidos. Assim, se nada for encontrado, uma mensagem será retornada ao usuário para alertar sobre o fato de nenhuma estrutura ter sido encontrada. Caso contrário, os PDBs recuperados serão apresentados ao usuário, conforme pode ser visto pela Figura 2.2.

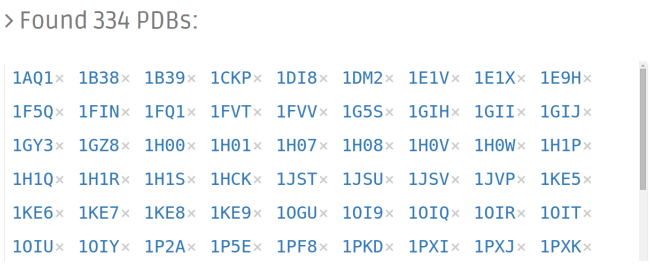


Figura 2.2. Resultado de uma pesquisa de PDBs.

Destacamos que o nAPOLI aplica um filtro automático que exclui de antemão todas as estruturas que não possuam ligantes. Além disso, vale ressaltar que o identificador PDB definido na pesquisa não necessariamente precisa ter um ligante em sua estrutura. Isto porque, caso o usuário esteja interessado em encontrar prováveis ligantes para uma certa estrutura, esse poderá primeiramente encontrar proteínas similares à proteína em estudo e investigar os padrões de interações que existem perante às proteínas similares encontradas e seus ligantes.

Além de poder ver cada arquivo recuperado, o usuário também tem a opção de remover os arquivos que não sejam de seu interesse, rever os parâmetros utilizados, clicar com o mouse sobre um identificador PDB para visitar a página referente a ele no site do PDB a fim de obter informações detalhadas sobre essa estrutura ou ainda definir os novos parâmetros que serão utilizados no próximo passo a ser descrito em seguida.

Após compor a base de dados a ser utilizada – o conjunto de identificadores PDB

recuperado – o usuário deverá, neste instante, definir alguns parâmetros que serão utilizados na próxima pesquisa: a pesquisa de grupos de ligantes.

2.4.2 Encontrando grupos de ligantes por regiões

Implementamos um algoritmo no nAPOLI que visa buscar grupos de ligantes que estejam localizados em uma mesma região. As regiões são definidas de acordo com um dos parâmetros do usuário conforme será mostrado posteriormente. A ideia dessa etapa se baseia no fato de que muitas vezes o pesquisador não possui o conhecimento de quais e onde os ligantes estão localizados em uma estrutura. Portanto, a ferramenta encontrará e fornecerá ao usuário estas informações que serão utilizadas para encontrar padrões de interações proteína-ligante. Primeiramente, serão apresentados todos os parâmetros disponíveis e uma breve descrição sobre cada um e finalmente será apresentado o funcionamento do algoritmo.

A Figura 2.3 a seguir demonstra os parâmetros disponíveis para a busca de grupos de ligantes. São eles:

- Limiar de distância (*Cutoff distance*) em Ångström: define o raio de distância utilizado para encontrar ligantes em uma região. Por padrão, o valor utilizado para esse campo é 5Å.
- Número mínimo de átomos (*Minimum number of atoms*): define um número mínimo de átomos pesados (átomos que não sejam átomos de hidrogênio) que um ligante deve ter para não ser considerado artefato. Por padrão, um ligante deve ter 7 ou mais átomos pesados para não ser considerado artefato de Cristalografia [Pires et al., 2013].
- Conjunto de artefatos de Cristalografia (*Set of Artifacts of Crystallography*): conjunto de identificadores referentes aos artefatos de cristalografia que o usuário deseja desconsiderar. Artefatos podem ser sais, soluções tampão e outros aditivos que são incluídos na técnica de cristalografia por difração de raios X para induzir a formação dos cristais [Forli, 2015; Rupp, 2009]. Por padrão, consideramos os seguintes ligantes como artefatos: ACE, ACT, BME, CSD, CSW, EDO, FMT, GOL, MSE, NAG, NO3, PO4, SGM, SO4, TPO [Bricogne et al., 2011; Strömbergsson e Kleywegt, 2009].

O limiar de distância padrão foi definido segundo curadorias manuais e empíricas que realizamos com alguns conjuntos de PDBs. Com o valor de 5Å, observamos que ligantes em uma região de sítio de ligação eram sempre adicionados no mesmo grupo,

assim, consideramos esse valor como ideal. À medida que esse valor é aumentado ligantes mais distantes e que estão longe do sítio de ligação também são adicionados ao mesmo grupo, que é algo não desejável uma vez que esses ligantes em grande parte eram artefatos de cristalografia. Valores inferiores a 5Å em certas ocasiões não foram hábeis em capturar todos os ligantes de uma região. Apesar de um teste experimental envolvendo todos os sítios de ligação disponíveis no site do PDB não ter sido feito, ressalta-se que esse parâmetro pode ser definido pelo usuário caso ele sinta a necessidade.

Figura 2.3. Painel com os parâmetros disponíveis para a busca de grupos de ligantes.

Um pré-processamento antecede as etapas principais e tem como objetivo filtrar os arquivos PDB a fim de preparar e gerar uma estrutura apropriada para o perfeito funcionamento do algoritmo. Um dos filtros mais importantes nessa etapa são os filtros que removem ácidos nucleicos e aminoácidos não-padrão, isto é, aminoácidos que não pertencem à lista de aminoácidos mais comumente encontrados em seres vivos (Tabela 1.1).

A remoção de ácidos nucleicos se deve ao fato de que todos os cálculos realizados no nAPOLI se baseiam em interações entre proteína e ligante. Além disso, o segundo motivo para se aplicar esse filtro é que o software (LovoAlign [Martínez et al., 2007]) que utilizamos no alinhamento de estruturas não permite ácidos nucleicos como entrada para o alinhamento.

A seguir, fazemos também a remoção de átomos de hidrogênio e moléculas de água, a escolha do primeiro modelo quando disponível na estrutura resolvida por ressonância magnética e filtro para posições alternativas, em que o átomo que possui a maior ocupância é escolhido, enquanto que o outro átomo é removido. Na Seção 2.5 há uma descrição sobre cada filtro aplicado.

Finalmente, o passo inicial da etapa de pesquisa de ligantes se dá a partir de um alinhamento estrutural utilizando o software de alinhamento, LovoAlign [Martínez et al., 2007] e para isso utiliza-se como modelo (*template*) o identificador PDB e a cadeia informados na pesquisa de PDBs (Seção 2.4.1). O modelo é uma estrutura que será

utilizada como base para o alinhamento, de forma que todas as demais estruturas serão alinhadas de acordo com a estrutura do modelo. LovoAlign é útil nesse processo porque cria novos arquivos PDB em que as coordenadas dos átomos de todas as estruturas alinhadas são alteradas para equivaler às coordenadas da estrutura modelo. Assim, todos os resíduos que são alinhados na mesma posição terão a mesma coordenada.

A ideia principal por trás dessa etapa é sobrepor as estruturas escolhidas pelo usuário de uma forma que as regiões contendo ligantes também fiquem sobrepostas. Por exemplo, suponha que a base de dados escolhida seja composta por diversos arquivos PDB no qual suas estruturas representem uma mesma proteína, porém, em complexo com um ligante diferente. Assim, a partir desse alinhamento, espera-se que as regiões alvo ou *pockets* – como são comumente conhecidas estas regiões [Kessel e Ben-Tal, 2010] – desses ligantes estarão sobrepostas; dessa forma, os ligantes localizados nessas regiões e que estão em identificadores PDB diferentes serão facilmente encontrados na etapa que se segue.

Após realizados todos os alinhamentos, a busca por regiões que contenham ligantes se inicia. Primeiramente, o algoritmo deve buscar todos os ligantes válidos que estejam disponíveis no arquivo PDB. Nesse ponto, consideramos válidos todos os ligantes que não sejam íons e para tanto, utilizamos como referência uma lista de íons (Tabela 2.1) disponíveis na página Hic-Up[11] [Kleywegt, 2007].

Tabela 2.1. Lista de identificadores dos ligantes (íons) não permitidos no nAPOLI.

Identificadores dos íons
3CO, 3NI, 4MO, 6MO, ARS, AU3, CH2, CH3, CU1, EU3, FE2, GD3, GTE, H2S, HO3, HYD, IOD, IR3, MN3, ND4, NH2, NH3, NH4, OS4, PC4, PT4, QTR, SEK, YB2, YT3, AG, AL, AR, AU, BA, BR, CA, CD, CE, CL, CO, CR, CS, CU, EU, FE, GA, GD, HG, HO, IN, IR, KR, LA, LI, LU, MG, MN, MO, NA, NI, OH, OS, OX, PB, PD, PR, PT, RB, RE, RU, SB, SE, SM, SR, TB, TE, TL, U1, XE, Y1, YB, ZN, F, K, O, S, V, W

Em seguida, o algoritmo tentará agrupar os ligantes de acordo com a região em que esses estão localizados, segundo as coordenadas de seus átomos. Para isso, utilizamos a distância euclidiana com uma abordagem todos contra todos a nível atômico. A distância euclidiana pode ser calculada segundo a equaçao abaixo:

$$D(i,j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \tag{2.1}$$

[11] http://xray.bmc.uu.se/hicup/

Em que, $D(i, j)$ é a distância entre os átomos i e j, e suas posições correspondentes no espaço tridimensional são definidas pelas coordenadas (x_i, y_i, z_i) e (x_j, y_j, z_j), respectivamente.

Assim, para cada par de ligantes e para cada par de átomos entre esses dois ligantes, calcula-se a distância euclidiana. Se qualquer um destes pares de átomos estiverem a uma distância menor que o limiar definido pelo usuário (*Cutoff distance*), agrupamos esses ligantes, o que indica que eles estão em uma mesma região. Além disso, o agrupamento é feito de forma transitiva, isto é, suponha que existam três ligantes A, B e C em uma certa região. A encontra-se próximo de B, mas distante de C. E por sua vez, C está próximo de B, mas distante de A. Como A e C são vizinhos de um mesmo ligante (B), eles serão agrupados em um mesmo grupo.

A Figura 2.4 ilustra o resultado desse processo de agrupamento. Em 2.4.*a* exemplifica-se de forma global os agrupamentos obtidos após o alinhamento. Nessa figura, cada ligante foi colorido de acordo com os átomos que o compõe utilizando a abordagem de cores conhecida como CPK – esse esquema de cores define uma cor para cada elemento químico de acordo com Corey [1953]. Além disso, exemplifica-se também nessa figura o número de ligantes que compõem três grupos escolhidos como exemplo. Em dois destes grupos (16 e 10 ligantes), cada ligante pertence a um identificador PDB diferente, enquanto que no terceiro grupo, 2 desses ligantes pertencem ao mesmo PDB.

Vale ressaltar que o agrupamento se dá em um plano tridimensional, assim, uma visão em apenas duas dimensões como a da figura pode induzir alguém a presumir que alguns grupos deveriam formar um único grupo devido à aparente proximidade desses agrupamentos.

Já em 2.4.*b*, a região destacada com um círculo vermelho na Figura 2.4.*b* é mostrada em detalhes juntamente com a visão em *cartoon* das estruturas alinhadas. Nessa figura, são mostrados 5 de 15 ligantes dessa região. Cada ligante foi colorido de uma cor diferente para facilitar a visualização e para poder ilustrar uma região contendo ligantes que pertencem a diferentes identificadores PDB.

Finalizado o agrupamento, a próxima etapa é a identificação de possíveis artefatos de cristalografia. Para isso, utiliza-se o valor definido em *Minimum number of atoms* e em *Set of artifacts of crystallography* que foram descritos previamente. Essa etapa tem como objetivo ajudar o usuário a identificar possíveis ligantes que não serão de seu interesse e os quais o usuário provavelmente não deseja utilizar na identificação de padrões de interações proteína-ligante.

O processo de identificação de artefatos se dá de duas formas. O primeiro filtro verifica se um determinado ligante está presente na lista de artefatos definida pelo usuário (*Set of artifacts of crystallography*), nesse caso, o algoritmo automaticamente

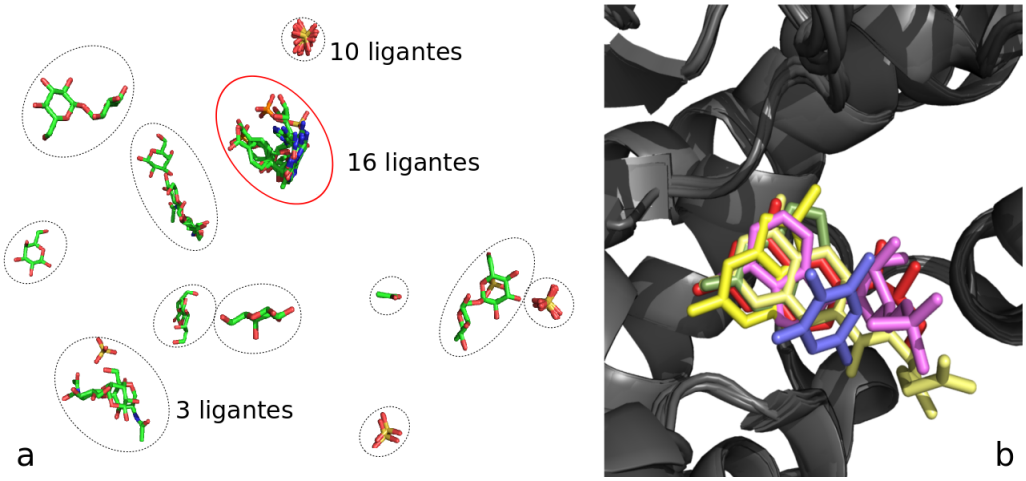


Figura 2.4. Resultado da busca de ligantes. Em (a), são mostrados os grupos de ligantes encontrados nas estruturas alinhadas. Junto de três destes grupos, é mostrado o número de ligantes encontrados naquela região. Em (b), mostra-se, em detalhes, 5 de 15 ligantes encontrados na região destacada em vermelho na Figura (a). As estruturas alinhadas são mostradas em *cartoon* e os ligantes em *stick*, sendo coloridos cada um de uma cor em (b).

considera o ligante como artefato. Caso não esteja presente na lista, o segundo filtro computará o número de átomos pesados presentes no ligante e sendo esse número menor que aquele definido em *Minimum number of atoms*, o algoritmo também irá considerar o ligante como artefato. Após a identificação desses ligantes, caberá ao usuário a decisão de utilizá-los ou removê-los de sua base de dados.

Por fim, o resultado obtido em todo esse processo será retornado ao usuário e ele terá acesso a um relatório descrevendo informações tais como: possíveis erros durante o alinhamento, lista de íons encontrados e que foram ignorados, uma lista de identificadores PDB que não possuem ligantes válidos, algumas estatísticas diversas sobre o número de ligantes encontrados e a lista de ligantes agrupados por regiões.

Tendo a lista de ligantes válidos agrupados por regiões, o usuário poderá escolher o grupo que mais lhe convier e submeter seus dados para a etapa final: a análise de padrões de interações. Entretanto, caso o usuário deseje, antes de realizar a submissão de seus dados, ele poderá ainda configurar alguns parâmetros que serão utilizados na próxima etapa.

Esses parâmetros definem um limiar de distância (em Ångströms) para calcular as interações entre dois átomos e serão melhor descritos na Seção 2.8. Neste momento, vale listar cada um destes parâmetros juntamente com seus valores padrão definidos de acordo com Mancini et al. [2004]:

- Empilhamento aromático (*Aromatic stacking*): de 1,5 a 2,8 Å;

- Interação hidrofóbica (*Hydrophobic interaction*): de 2,0 a 3,8 Å;
- Interação repulsiva *Repulsive interaction*: de 2,0 a 6,0 Å;
- Pontes de hidrogênio (*Hydrogen bond*): de 2,0 a 3,2 Å;
- Ponte salina ou interação atrativa (*Salt bridge*): de 2,0 a 6,0 Å.

2.5 Download e filtragem de PDBs

Após o inicio da busca de padrões de um conjunto de PDBs e ligantes, o nAPOLI primeiramente verificará se cada arquivo PDB definido pelo usuário já se encontra disponível na base de dados interna da ferramenta. Caso o PDB não esteja disponível, deve-se realizar o *download* desse PDB a partir do banco de dados de estruturas RCSB PDB. O *download* é feito de forma automática utilizando o seguinte endereço: http://www.rcsb.org/pdb/files/PDB. Note que nesse endereço há um campo de nome PDB, esse campo será substituído pelo identificador correspondente ao PDB a ser obtido dessa base de dados. O nAPOLI implementa um sistema de processamento paralelo para realizar os *downloads* de forma simultânea de acordo com o número de processadores disponíveis.

Entretanto, esse método de checagem de PDBs ainda não disponíveis é um método que pode ser considerado preventivo. Isto porque nAPOLI dispõe de um algoritmo que realiza a sincronização de todos os arquivos PDBs disponíveis em RCSB PDB semanalmente a fim de manter a base de dados do nAPOLI sempre atualizada. Porém, como a sincronização é realizada apenas uma vez por semana, há a possibilidade de que a base de dados do nAPOLI fique desatualizada devido a novas estruturas que por ventura sejam depositadas no RCSB PDB antes que uma nova sincronização seja feita. Portanto, para evitar que um usuário tente utilizar um PDB que ainda não esteja disponível no nAPOLI, verifica-se sempre a existência dos PDBs selecionados pelo usuário.

Após o *download* de cada PDB, é feita uma filtragem nestes arquivos com o objetivo de mantê-los no padrão necessário para o funcionamento pleno do nAPOLI. Em geral, os seguintes filtros são realizados: remoção de hidrogênio e moléculas de água; remoção de resíduos que não sejam algum dos 20 aminoácidos mais comuns (Tabela 1.1); caso exista mais de um modelo disponível no arquivo PDB, apenas o primeiro modelo é mantido; caso exista posições alternativas (átomos com ocupância), o átomo com a maior ocupância é mantido.

Neste trabalho, optamos pela remoção de hidrogênios primeiramente como uma forma de padronização, uma vez que algumas estruturas depositadas no RCSB PDB possuem hidrogênio em seus resíduos e ligantes; por outro lado, diversas estruturas não apresentam hidrogênios na composição de resíduos e ligantes. Além disso, o segundo motivo se deve ao fato de que os métodos aqui utilizados para encontrar interações entre átomos não definem interações utilizando esse átomo, mesmo que ele esteja envolvido em uma ponte de hidrogênio. Nesse caso, a interação é mapeada considerando diretamente os dois outros átomos envolvidos nessa interação, isto é, o átomo aceptor e o átomo doador.

Moléculas de água (H_2O) e aminoácidos não-padrão – incluindo nessa classe ácidos nucleicos – foram desconsiderados como uma forma de simplificação do problema. Sabe-se que diversas moléculas de água participam diretamente no processo de reconhecimento molecular [Davey et al., 2002; Ernst et al., 1995; Jayaram e Jain, 2004; Levy e Onuchic, 2006; Meyer, 1992; Park e Saven, 2005; Schwabe, 1997; Sreenivasan e Axelsen, 1992], e desconsiderá-las pode ocasionar a perda de informações importantes relacionadas às interações entre uma proteína e seu ligante. Assim sendo, em uma próxima versão iremos incluir moléculas de água em nossa metodologia.

Com relação aos aminoácidos não-padrão, isto é, resíduos de aminoácidos que não pertençam à lista dos 20 aminoácidos mais comumente encontrados nos seres vivos, optou-se por ignorá-los neste trabalho também por fins de simplificação do problema. Definiu-se a partir dos trabalhos de [de Melo et al., 2006, 2007b,a; Pires et al., 2011; da Silveira et al., 2009], o tipo de cada átomo de todos os 20 aminoácidos mais comuns utilizando-se as propriedades químicas de cada átomo (Seção 2.7.1); porém, esse mapeamento foi realizado apenas para estes aminoácidos considerando, além disso, um ambiente com pH 7. Assim, dado o mapeamento disponível, qualquer resíduo que não tenha sido mapeado de acordo com as propriedades de seus átomos não poderá ser utilizado na versão atual da ferramenta.

Em geral, técnicas como Ressonância Magnética Nuclear (RMN) produzem vários modelos para representar uma estrutura [Kessel e Ben-Tal, 2010]. Portanto, neste trabalho decidimos utilizar apenas o primeiro modelo disponível em um arquivo PDB para fins de simplificação do problema. Em uma próxima versão da ferramenta, podemos oferecer meios para que o usuário análise padrões de interação proteína-ligante utilizando cada modelo disponível em um arquivo PDB.

Por outro lado, a cristalografia por difração de raios X utiliza uma abordagem de resolução de estruturas de proteínas a partir da formação de cristais. Cada cristal obtido é formado por diversas proteínas todas de um mesmo tipo. Após todo o processo de obtenção da estrutura, é possível visualizar o posicionamento dos átomos de cada

aminoácido pertencente à proteína em estudo. Porém, como cada cristal apresenta diversas proteínas não se pode assegurar que todas as proteínas estarão em uma mesma conformação. Com isso, ao se comparar um resíduo na mesma posição da sequência de aminoácidos em cada proteína cristalizada, podem ser encontrados átomos que estejam em posições ligeiramente diferentes na estrutura [Verli, 2014]. Por isso, alguns PDBs possuem a informação de todas as posições que um átomo foi encontrado. Assim, optamos neste trabalho por considerar apenas as posições que foram observadas mais vezes no conjunto de estruturas resolvidas, isto, aquelas que apresentam o maior valor de ocupância.

2.6 Mapeamento de contatos entre átomos

Um passo que precede o cálculo das interações e obrigatório, é mapear quais átomos estão em contato. Segundo da Silveira et al. [2009], contato é um termo que se refere apenas à presença e distribuição espacial dos átomos, sendo útil para definir quais átomos estão próximos uns dos outros. Enquanto que as interações se referem a toda ação entre dois átomos quaisquer em resposta a algum tipo de força mútua como, por exemplo, as interações decorrentes de uma repulsão entre os átomos.

Segundo da Silveira et al. [2009] as duas principais abordagens para calcular contato são: dependente de corte (*cutoff dependent*) e a independente de corte (*cutoff free*).

A abordagem dependente de corte, como o próprio nome diz, exige que seja definido um valor a ser utilizado como limiar de distância tal como ilustrado pela Figura 2.5. Nessa abordagem, os contatos são definidos como se segue: seja uma esfera (círculo tracejado em vermelho) de raio r (*cutoff*) centrada em um átomo X (em vermelho); qualquer outro átomo (círculos menores) que estiver localizado dentro da esfera definida pelo raio r estará em contato com o átomo X. Um primeiro problema que surge com essa abordagem se refere à definição do valor de corte que não é algo trivial de ser feito [da Silveira et al., 2009].

Outro problema frequente são os contatos oclusos (falsos contatos), também representado na Figura 2.5. Eles ocorrem quando dois átomos X e Y (em azul) são mapeados como átomos em contato, mesmo que haja um terceiro átomo (em amarelo) entre eles intervindo. Uma interação eletrostática, por exemplo, entre os átomos X e Y não ocorrerá uma vez que há um átomo entre eles intervindo a interação [Dokholyan, 2012].

Uma forma de evitar estes problemas é a partir do uso de uma abordagem in-

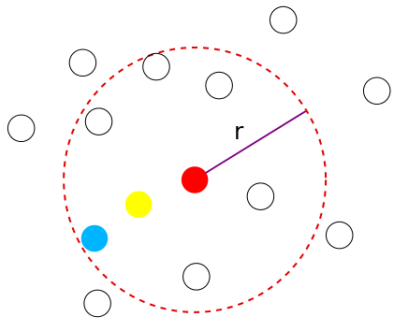


Figura 2.5. Abordagem dependente de corte. Círculos menores são átomos e a esfera (círculo tracejado em vermelho) de raio r (linha em lilás) centrada em um átomo X (em vermelho) define quais átomos estão em contato com o átomo X. O contato entre o átomo X e um átomo Y (em azul) é considerado um contato ocluso porque há outro átomo (em amarelo) entre eles intervindo a interação.

dependente de valor de corte. Além de não ser necessário definir esse valor, há uma garantia matemática de que essa abordagem é livre de oclusão [da Silveira et al., 2009]. Portanto, neste trabalho decidimos utilizar essa técnica para evitar falsos positivos (contatos oclusos). Pretendemos, porém, em uma próxima versão, oferecer ambas as opções para que o usuário decida a abordagem de sua preferência.

Uma forma de aplicar a abordagem independente de corte é através do uso de diagramas de Voronoi e seu grafo dual, a triangulação de Delaunay. Diagramas de Voronoi, em homenagem ao matemático russo Georgy Voronoi, é uma técnica que aplica conceitos puramente geométricos [Poupon, 2004] e consistem basicamente em dividir o espaço em regiões, chamadas de células de Voronoi, de acordo com o número de átomos existentes nesse espaço. Sendo que cada átomo pertencerá sempre a uma única célula.

A construção de uma célula de Voronoi consiste em traçar os planos que cortam as linhas desenhadas a partir de um centroide para todos os outros centroides e definir a célula como sendo o menor poliedro formado por esses planos (Figura 2.6(a)). No caso da representação de um diagrama de Voronoi para uma proteína, os átomos serão os centroides e são, então, utilizados para gerar as células de Voronoi (Figura 2.6(b)). O conjunto de células preenchem o espaço e definem uma tesselação. Em geometria, uma tesselação é o preenchimento do espaço utilizando-se uma forma geométrica, de modo que o espaço seja totalmente preenchido, sem espaços e lacunas.

A triangulação de Delaunay pode ser obtida a partir desse diagrama gerado. Para isso, basta que sejam traçados arestas entre todos os centroides (átomos nesse caso) que tenham uma face comum em suas células de Voronoi, ou seja, qualquer célula que

esteja em contato direto com outra célula (Figura 2.6(c)) [Poupon, 2004]. Dessa forma, apenas os átomos mais próximos e não oclusos estarão conectados e fazendo contato um com o outro [Gonçalves-Almeida, 2011].

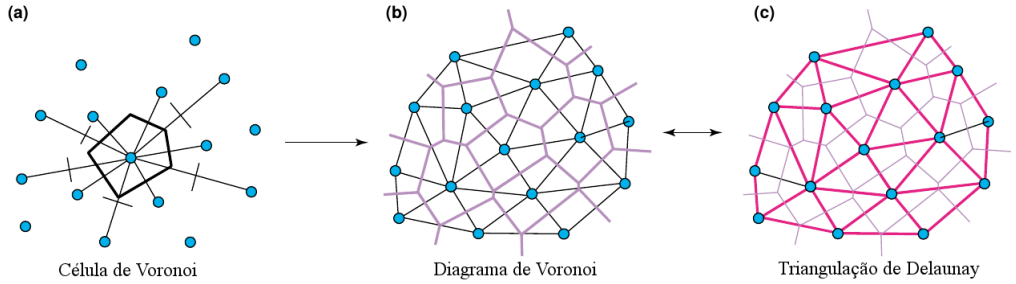


Figura 2.6. Diagramas de Voronoi e triangulação de Delaunay. Em (a), um exemplo de como são geradas as células de Voronoi. Em (b), um diagrama de Voronoi (em lilás) e as arestas (linhas em preto) conectando os centroides que possuem uma face em comum. Em (c), uma triangulação de Delaunay (em rosa) e um diagrama de Voronoi (em lilás). Adaptado de Poupon [2004].

De modo inverso, também é possível gerar um diagrama de Voronoi a partir de uma triangulação de Delaunay, uma vez que eles são duais. Além disso, para cada diagrama de Voronoi existe apenas uma triangulação de Delaunay [Poupon, 2004].

Como resultado, após conectar todos os pares de centroides em que suas células de Voronoi estão em contato, será formado um grafo onde os átomos são os vértices e as arestas são os contatos entre os átomos. Esses contatos serão posteriormente utilizados para mapear as interações. Na Seção 2.13, serão abordados alguns conceitos sobre grafos.

O diagrama de Voronoi e seu dual, a triangulação de Delaunay, foram computados utilizando o *software* CGAL [1995], que é uma biblioteca escrita em C++ e que oferece uma série de algoritmos para cálculos geométricos.

Neste trabalho, estamos interessados apenas nas interações entre a proteína e o ligante, isto é, a interface da proteína com o ligante. Portanto, contatos formados por dois átomos da proteína ou entre dois átomos do ligante são descartados.

2.7 Tipos de átomos

As interações e a forma como estas ocorrem no macroambiente celular dependem de diversos fatores físicos, químicos e biológicos. Primeiramente, as distâncias entre dois átomos talvez sejam as mais facilmente compreensíveis devido às propriedades físicas. Por exemplo, a atração e repulsão entre dois átomos carregados define que

forças de atração e repulsão são inversamente proporcionais ao quadrado da distância entre dois átomos carregados (Lei de Coulomb) tanto positivamente (cátions) quanto negativamente (ânions).

Entretanto, muitas outras forças e propriedades estão relacionadas a esse processo o que, portanto, explica a complexidade envolvendo interações e reconhecimentos moleculares entre duas ou mais moléculas. Solvatação, polaridade de grupos químicos, pH do ambiente, se há íons participando da interação, hidrofobicidade e tipos dos átomos envolvidos na interação são passíveis de serem utilizados como objeto de estudo para explicar como ocorrem as interações entre moléculas. Porém, uma modelagem que é mais comumente utilizada é aquela em que são consideradas apenas os tipos de átomos e a distância que há entre eles. De certa forma, classificar um átomo de acordo com uma propriedade química também inclui adicionar informações como hidrofobicidade e polaridade presente nos átomos.

Assim, nesye trabalho utilizamos o mapeamento de interações utilizando critérios de distância e os tipos de cada átomo envolvido na interação. Ressalta-se que o cálculo de interações depende também dos contatos entre átomos gerados conforme descrito na Seção 2.6.

Dessa forma, as interações somente podem ser mapeadas caso os tipos de átomos estejam disponíveis. Para tanto, os átomos compondo aminoácidos e ligantes foram classificados pelas suas propriedades físico-químicas de acordo com o trabalho de Sobolev et al. [1999]. Os átomos foram classificados como sendo: aceptor, aromático, doador, hidrofóbico, negativo (ânion) ou positivo (cátion). Nessa abordagem, alguns átomos podem ser classificados com mais de uma propriedade. Por exemplo, o nitrogênio NE2 de uma histidina foi classificado como sendo aromático, doador e positivo.

A seguir serão descritas as metodologias utilizadas para classificar cada tipo de átomo de acordo com a molécula ao qual um átomo pertence.

2.7.1 Átomos do aminoácido

A classificação de cada átomo de um aminoácido foi feito manualmente em nossos trabalhos anteriores considerando um pH igual a 7 [de Melo et al., 2006, 2007b,a; Pires et al., 2011; da Silveira et al., 2009]. Nessa classificação considerou-se os vinte aminoácidos mais comuns (Tabela 1.1) e as propriedades físico-químicas dos átomos de acordo com o trabalho de Sobolev et al. [1999].

Tabela 2.2. Classificação de cada átomo de um aminoácido segundo suas propriedades físico-químicas.

Tipo de átomo	Átomos classificados em um certo tipo
Aceptores	ALA (O), ARG (O), ASN (O, OD1), ASP (O, OD1, OD2), CYS (O), GLN (O, OE1), GLU (O, OE, OE2), GLY (O), HIS (O), ILE (O), LEU (O), LYS (O), MET (O), PHE (O), PRO (O), SER (O), THR (O), 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, OD1), ASP (N), CYS (N), GLN (N), GLU (N), GLY (N), HIS (N, ND1, NE2), ILE (N), LEU (N), LYS (N, NZ), MET (N), PHE (N), PRO (N), SER (N, OG), THR (N, OG1), TRP (N, NE1), TYR (N, OH), VAL (N)
Hidrofóbicos	ALA (CB), ARG (CB, CD, CG, CZ), ASN (CB, CG), ASP (CB, CG), CYS (CB), GLN (CB, CD, CG), GLU (CB, CD, CG), HIS (CB, CD2, CE1, CG), ILE (CB, CD1, CG1, CG2), LEU (CB, CD1, CD2, CG), LYS (CB, CD, CE, CG), MET (CB, CE, CG, SD), PHE (CB, CD1, CD2, CE1, CE2, CG, CZ), PRO (CB, CD, CG), SER (CB), THR (CB, CG2), TRP (CB, CD1, CD2, CE2, CE3, CG, CH2, CZ2, CZ3), TYR (CB, CD1, CD2, CE1, CE2, CG, CZ), VAL (CB, CG1, CG2)
Negativos	ASP (OD1, OD2), GLU (OE1, OE2)
Positivos	ARG (NH1, NH2), HIS (ND1, NE2) e LYS (NZ)

2.7.2 Átomos do ligante

O mapeamento para os átomos do ligante foi realizado de acordo com os mesmos tipos de átomos utilizados para os aminoácidos, isto é, os átomos foram classificados como aceptor, aromático, doador, hidrofóbico, negativo (ânion) ou positivo (cátion) conforme o trabalho de Sobolev et al. [1999].

Conforme especificado na Seção 2.7.1, o mapeamento manual foi realizado apenas para aminoácidos. Dessa forma, para mapear os tipos de átomos para o ligante utilizamos o *software* Pmapper (GenerateMD 5.3.8, 2010, ChemAxon [12]). Além disso, conforme foi feito para os aminoácidos, o mapeamento para os átomos do ligante também foi realizado considerando um pH igual a 7. O Pmapper tem a capacidade de encontrar propriedades farmacofóricas de átomos para uma estrutura molecular de acordo com diversas propriedades químicas e estruturais como carga e hidrofobicidade [Chemaxon, 2014].

[12] http://www.chemaxon.com

Previamente, deve-se realizar um pré-processamento dos arquivos contendo a estrutura dos ligantes. O primeiro passo é extrair do arquivo PDB todas as informações dos átomos do ligante alvo e também as informações de outras moléculas que estejam ligados covalentemente a essa molécula alvo. Isso é importante porque essas ligações covalentes podem influenciar no resultado obtido pelo Pmapper.

Para ilustrar esse procedimento, tomemos como o exemplo o ligante BMA1455 encontrado na cadeia A do PDB 1H4P. Esse ligante, apesar de não estar localizado em um sítio de ligação para essa proteína, e apesar de não termos encontrado interação direta dele com a proteína, ainda assim pode ser utilizado para ilustrar o efeito que as ligações covalentes podem gerar no resultado obtido.

A Figura 2.7(a) mostra o ligante BMA1455 ao centro da imagem colorido de acordo com o esquema de cores CPK e os demais ligantes que se ligam covalentemente a ele em amarelo. Os átomos do ligante BMA1455 que estão ligados covalentemente são mostrados em esferas (O3, O6 e C1). Quando esse ligante está isolado (sem nenhuma outra molécula ligada a ele)(Figura 2.7(b)) os dois átomos de oxigênio estão ligados a um hidrogênio, o que faz com que estes átomos sejam classificados pelo Pmapper como aceptor e doador de hidrogênios. Porém, quando estes oxigênios se ligam covalentemente a outra molécula, como mostrado em (a), não há mais a presença desses hidrogênios, então, eles passam a ser classificados apenas como aceptores. Por sua vez, o oxigênio O4 (também representado por uma esfera em (a)), é classificado como aceptor e doador em ambos os casos, pois não apresenta ligações covalentes com outras moléculas. Quanto ao carbono C1, em ambos os casos é classificado como hidrofóbico. Todas estas constatações estão de acordo com o trabalho de Sobolev et al. [1999].

Vale ressaltar que, nesse tipo de situação, as ligações covalentes existem com a finalidade de unir todos os ligantes e, assim, formar um único ligante maior, isto é, um ligante formado por diversas componentes (ligantes). Por exemplo, os dissacarídeos podem ser considerados ligantes maiores formados a partir de ligações covalentes (ligações glicosídicas) entre dois monossacarídeos [Nelson e Cox, 2014]. Assim, ligantes covalentemente ligados devem ser analisados como se fossem um único ligante tal como é feito por de Beer et al. [2014]. Contudo, na atual versão do nAPOLI consideramos as ligações covalentes apenas durante a classificação dos átomos dos ligantes. A análise das interações proteína-ligante, porém, é feita para cada uma das componentes separadamente. Portanto, em versões futuras planejamos realizar essas análises considerando ligantes covalentemente ligados como se fossem um único ligante.

Assim, após extrair as informações dos átomos, utiliza-se o *software* Open Babel [O'Boyle et al., 2011; Open Babel, 2012] para converter o arquivo do formato PDB para o formato Molfile (.mol). Finalmente, os arquivos estão prontos para que o Pmapper

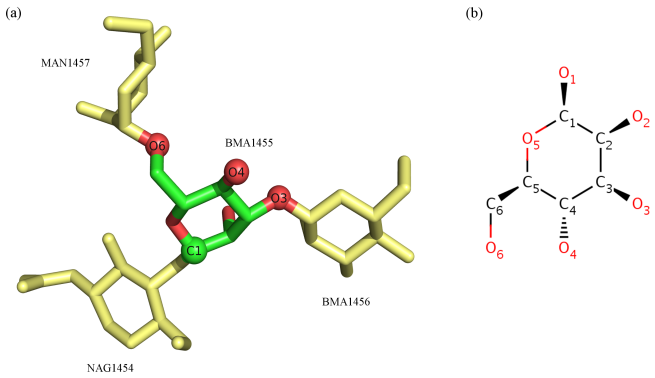


Figura 2.7. Efeito de ligações covalentes na classificação dos tipos de átomo. Em (a), os átomos O3 e O6 do ligante BMA1455 são classificados como aceptores devido às ligações covalentes entre eles e dois outros ligantes. Quando o BMA1455 está isolado (b), estes mesmos átomos são classificados como aceptores e doadores.

possa caracterizar cada átomo quanto aos seus tipos. A conversão do formato PDB para o formato Molfile se faz necessária, pois conforme a própria Chemaxon[13], arquivos PDBs podem ser analisados de forma incorreta.

2.8 Mapeamento de interações entre átomos

O mapeamento de interações foi feito a partir de três informações: contatos, distância entre os átomos em contato e o tipo de cada átomo. Os contatos foram gerados tal como descrito na Seção 2.6 e os tipos de átomos foram mapeados como descrito na Seção 2.7. Quanto à distância entre dois átomos, é calculada a partir da fórmula da distância euclidiana (Equação 2.1). Dessa forma, pode-se dizer que o algoritmo aqui utilizado leva em consideração propriedades físico-químicas e as distâncias entre os átomos.

Basicamente, dois átomos estarão interagindo se e apenas se a distância euclidiana entre os dois estiver dentro de um intervalo previamente definido pelo usuário e se os tipos dos átomos favorecerem o estabelecimento de algum tipo de interação. A tabela a seguir descreve as combinações de tipos e as interações possíveis para cada caso.

O intervalo para cada tipo de interação pode ser definido isoladamente pelo usuário como descrito na Seção 2.4.2. Porém, o usuário pode optar por não definir os valores para esses intervalos. Assim, serão utilizados os valores padrões:

- Empilhamento aromático: de 1,5 a 2,8Å;

[13]https://docs.chemaxon.com/display/FF/Protein+Data+Bank+%28PDB%29+file+format

Tabela 2.3. Tipos de interações possíveis para as combinações de tipos atômicos.

Tipo do átomo 1	Tipo do átomo 2	Interação
Aceptor	Doador	Ponte de hidrogênio
Doador	Aceptor	Ponte de hidrogênio
Aromático	Aromático	Empilhamento aromático
Hidrofóbico	Hidrofóbico	Interação hidrofóbica
Positivo	Positivo	Interação repulsiva
Negativo	Negativo	Interação repulsiva
Positivo	Negativo	Ponte salina
Negativo	Positivo	Ponte salina

- Interação hidrofóbica: de 2 a 3,8Å;
- Interação repulsiva: de 2 a 6Å;
- Ponte de hidrogênio: de 2 a 3,2Å;
- Ponte salina: de 2 a 6Å.

Essas distâncias padrões foram definidas a partir de Mancini et al. [2004]. Por sua vez, esses autores utilizam a definição de contatos proposta por Sobolev et al. [1999].

Para ilustrar como o cálculo de interações ocorre, serão utilizadas as distâncias padrões aqui definidas. Primeiro caso: suponha que a distância euclidiana entre dois átomos A e B seja 2.8Å e os dois átomos sejam do tipo aromático. Então, a interação entre esses dois átomos será mapeada como Empilhamento aromático por causa dos tipos dos mesmos e porque a distância entre eles pertence ao intervalo [2;3,8]. Segundo caso: suponha que a distância euclidiana entre os átomos A e B seja 3.7Å e que estes átomos são do tipo aromático. Agora, não haverá nenhuma interação mapeada entre esses átomos uma vez que a distância está fora do intervalo [2;3,8].

Em certas circunstâncias, um átomo pode ser rotulado em mais de um tipo, como é o caso do átomo OE2 do aminoácido Glutamato que é rotulado como aceptor e negativo (Tabela 2.2). Esse átomo, então, pode interagir através de Ponte de hidrogênio, Ponte Salina ou Interação Repulsiva. Em nossa abordagem, definimos que um átomo irá interagir com outro átomo apenas através de um único tipo de interação. Para tanto, priorizamos as interações de acordo com a frequência com que elas ocorrem, tal como realizado por de Melo et al. [2006]. A lista de prioridade (em ordem decrescente) utilizada é:

- Ponte salina;
- Interação repulsiva;

- Empilhamento aromático;
- Ponte de hidrogênio;
- Interação hidrofóbica.

2.9 Gerando grupos de ligantes similares

Em um conjunto de ligantes é esperado encontrar uma grande variedade de moléculas constituídas por diferentes grupos químicos e propriedades físico-químicas. Dessa forma, um padrão global pode não emergir desse conjunto. Portanto, a análise de agrupamentos contendo ligantes similares se mostra necessária. A análise de cada agrupamento permite entender e descobrir padrões intrínsecos de cada agrupamento, bem como estudar os tipos de interações que cada agrupamento tende a estabelecer dado suas características próprias. Assim sendo, decidimos agrupar o conjunto de ligantes para que os usuários tenham a opção de analisar os padrões de interação proteína-ligante para cada grupo.

O primeiro passo a ser tomado é converter os ligantes para o formato .mol utilizando o *software* Open Babel [O'Boyle et al., 2011; Open Babel, 2012]. O formato .mol ou MDL Molfile é um formato para descrever moléculas criado pela empresa MDL Information Systems (MDL). Uma especificação detalhada desse formato pode ser encontrada em Dalby et al. [1992].

O próximo passo é criar *fingerprints* para cada ligante utilizando o *software* GenerateMD (GenerateMD 5.3.8, 2010, ChemAxon [14]). A ideia é transformar uma molécula em um descritor molecular que possa ser comparado e avaliado quanto à sua similaridade com outras moléculas. Segundo Todeschini et al. [2008], um descritor molecular é uma representação simbólica de informações químicas. Esses descritores podem conter informações tais como fórmulas químicas, peso molecular, dimensão molecular, polaridade, hidrofobicidade, pontos de fusão e ebulição, número de ligações rotacionais, entre outras Robson e Vaithiligam [2009].

2.9.1 *Fingerprints*

Fingerprint é um tipo de descritor molecular que caracteriza uma molécula bem como suas propriedades através de uma sequência de bits, isto é, uma sequência de zeros (0) e uns (1). Como cada *fingerprint* apresenta informações únicas de cada

[14] http://www.chemaxon.com

molécula, eles são muito utilizados para análise de similaridade entre as mesmas [Robson e Vaithiligam, 2009], que é justamente o que desejamos nessa etapa da ferramenta.

Existem duas abordagens principais para se definir um *fingerprint*: através de um *fingerprint* estrutural ou através de um *fingerprint* baseado em fragmentos (*hash fingerprint*). Para essa última abordagem, o *fingerprint* mais conhecido foi proposto pela Daylight [2011]. Assim, como o *software* aqui utilizado gera *fingerprints* de acordo com o modelo proposto pela Daylight, a descrição de *fingerprints* baseados em fragmentos será feita seguindo esse modelo.

Em um *fingerprint* estrutural, cada posição da sequência indica a presença ou ausência de uma estrutura previamente definida. Por exemplo, o primeiro bit da sequência pode indicar a presença (bit igual a 1) ou ausência (bit igual a 0) de um anel aromático; já a segunda posição poderia indicar a presença de grupo funcional amina e assim por diante. Dessa forma, se alguém estiver interessado em moléculas que tenham um anel aromático basta verificar se o primeiro bit de cada *fingerprint* é igual a 1 [Robson e Vaithiligam, 2009]. Porém, o tamanho da sequência de bits e a escolha de quais estruturas serão utilizadas para definir a sequência influencia diretamente no desempenho dos algoritmos de busca por similaridade. Assim, o padrão de estruturas a serem utilizadas depende muito do conjunto de ligantes da base de dados [Leach e Gillet, 2007].

Por outro lado, um *fingerprint* baseado em fragmentos não requer nenhum tipo de padrão de estruturas prévio sendo, portanto, mais genérico. Nesse contexto, uma posição da sequência não indica a presença ou ausência de nenhuma estrutura. Isto porque o *fingerprint* é gerado a partir da própria molécula, ou seja, somente as informações presentes nessa molécula definirão o *fingerprint*.

Por ser uma abordagem baseada em fragmentos, o algoritmo gera as seguintes informações:

- Padrões de tamanho um: padrões contendo apenas um átomo;
- Padrões de tamanho dois: cada par de átomos e a ligação entre eles;
- Padrões de tamanho três: cada trio de átomos e as ligações entre eles (duas ligações);
- Padrões de tamanho quatro: cada conjunto de quatro átomos e as ligações entre eles (três ligações);
- Padrões de tamanho N: cada conjunto de N átomos e a ligação entre eles (N - 1 ligação).

Em seguida, um algoritmo transforma cada um destes padrões em uma sequência que é utilizada para definir 1s em um certo número de bits do *fingerprint*, podendo ocorrer sobreposições nas posições da sequência de bits. Uma sobreposição ocorre quando dois padrões definem um bit na mesma posição do *fingerprint*. Por isso, uma posição na sequência de bits não reflete a presença ou ausência de uma estrutura.

Por outro lado, um padrão irá sempre definir os mesmos bits no *fingerprint*. Dessa forma, desde que pelo menos um bit dessa sequência gerada pelo algoritmo seja único (não compartilhado por nenhum outro padrão), pode-se dizer se um padrão está presente ou não. Quando um *fingerprint* indicar que um padrão não está presente, ele certamente não está. Apesar de um *fingerprint* não indicar com 100% de certeza que um padrão está presente, ele pode conter muito mais padrões no total do que um *fingerprint* estrutural. Segundo Daylight [2011], o resultado final de buscas por padrões e comparações entre *fingerprints* utilizando a abordagem baseada em fragmentos funciona muito melhor do que a abordagem estrutural.

Como mencionado anteriormente, utilizamos o *software* GenerateMD para gerar os *fingerprints* baseados em fragmento. Esse *software* recebe 5 parâmetros principais:

- Tamanho do *fingerprint*: define o número de bits que um *fingerprint* terá;
- Tamanho do padrão: define o número máximo de tamanhos de fragmentos a serem gerados, em que um tamanho igual a N irá gerar no máximo um padrão contendo N átomos e N-1 ligações;
- Número de bits para cada padrão: define o número de bits iguais a 1 que cada padrão (após ser transformado em uma sequência de bits) define no *fingerprint*;
- Tipo de descritor molecular: define se é um descritor químico ou se é um *fingerprint* de reações químicas ou se é um *fingerprint* descrevendo farmacóforos, etc;
- Formato da saída: define o formato dos *fingerprints* a ser salvo no arquivo de saída.

Os valores definidos para os parâmetros tamanho do *fingerprint*, tamanho do padrão e número de bits para cada padrão foram definidos conforme a recomendação dada pela própria Chemaxon[15]. O tamanho do *fingerprint* foi definido como sendo 1024, o tamanho do padrão como sendo 7 e número de bits para cada padrão como sendo 2.

[15]https://docs.chemaxon.com/display/CD/Chemical+Hashed+Fingerprint

Quanto ao tipo de descritor molecular utilizamos o valor 'CF' que define um descritor químico. Esse tipo de descritor gera *fingerprints* de acordo com as propriedades topológicas de cada molécula.

Finalmente, o formato da saída utilizado é o *decimal-format*. Esse formato define que os *fingerprints* devem ser transformados de valores binários (1s e 0s) para o formato decimal e define também que cada decimal deve ser separado por uma tabulação. Esse procedimento é feito porque o algoritmo que agrupa as moléculas de acordo com a similaridade exige que o formato do arquivo de entrada seja esse. Ressalta-se, porém, que a transformação feita não altera nenhum dos padrões que foram mapeados.

Tendo gerado todos os *fingerprints*, o próximo passo é o agrupamento por similaridade dos ligantes.

2.9.2 Agrupamento de ligantes por similaridade

O passo final para gerar os agrupamentos, é executar o *software* Ward (Ward 5.3.8, 2010, ChemAxon [16]) com os *fingerprints* gerados na etapa anterior.

Esse software utiliza o método proposto por Ward [1963], que é um método de agrupamento do tipo hierárquico aglomerativo. Essa classe de algoritmos inicia o agrupamento com N grupos, onde N é o número de elementos a serem agrupados – em nosso problema cada elemento representa um *fingeprint*. Nesse estágio inicial, cada grupo contém um único elemento (*singleton*). Em cada passo do algoritmo, é verificado se a união de dois grupos satisfaz uma determinada função objetivo. Isto procede, se desejado, até que haja apenas um único grupo formado por todos os elementos. No Ward, a função objetivo é dada pela soma dos quadrados dos erros [Downs e Barnard, 2003]. O agrupamento é sempre feito de forma a obter a menor soma dos quadrados dos erros, que é descrita como:

$$EES = \sum_{i=1}^{n} x_i^2 - \frac{1}{n} * (\sum_{i=1}^{n} x_i)^2 \tag{2.2}$$

Onde x_i representa o i-ésimo elemento e n é o número de elementos de um grupo. Se há m grupos, o EES total será a soma do EES individual de cada grupo.

Outro ponto importante em métodos de agrupamento hierárquico se refere ao número de agrupamentos desejado. Em geral, o número de agrupamentos possíveis para um determinado conjunto de elementos varia de 1 (um único grupo contendo todos os elementos) a N (um grupo para cada elemento). Assim, um algoritmo hierárquico pode agrupar os elementos em qualquer um desses valores disponíveis, que são chamados de

[16] http://www.chemaxon.com

níveis hierárquicos. Por exemplo, se o número de grupos desejado for 4, o algoritmo irá distribuir os elementos nesses 4 grupos de modo a obter a menor soma dos quadrados dos erros (no caso do Ward). Ressalta-se aqui que os agrupamentos são feitos de modo que um grupo seja formado por elementos similares entre si. Porém, em muitos casos, definir o número de agrupamentos no qual os elementos serão distribuídos de forma ótima não é algo trivial de ser feito [Downs e Barnard, 2003].

Para tanto, o *software* Ward disponibiliza uma opção que calcula e encontra o nível hierárquico ótimo a partir do uso do método proposto por Kelley et al. [1996]. A ideia desse método é basicamente encontrar o nível hierárquico no qual os elementos são melhor distribuídos, isto é, elementos similares irão pertencer a um mesmo grupo.

Assim, o primeiro passo é utilizar o *software* Ward com o método de Kelley [Kelley et al., 1996] para obter o número de agrupamentos ótimo. Em seguida, executa-se o Ward novamente, porém, utilizando o número ótimo de grupos obtido com o método de Kelley. Em ambos os casos, o parâmetro tamanho do *fingerprint* também deve ser definido (Seção 2.9.1). Nesse caso, utiliza-se o mesmo valor (1.024) definido para gerar os *fingerprints*. Ao final, o resultado destes passos será a formação de grupos de ligantes similares.

2.10 Estatísticas gerais

Nesta etapa, a ferramenta irá calcular estatísticas diversas relacionadas principalmente aos tipos atômicos e cada tipo de interação possível de ser estabelecida entre dois átomos. Essas e outras informações estarão dispostas em um formato de tabela onde cada coluna representa uma informação relevante às análises de padrões de interação proteína-ligante e cada linha representa a informação corresponde a uma estrutura PDB em específico. Essa tabela posteriormente será utilizada para produzir a frequência de cada tipo de átomo na base de dados como um todo, bem como a frequência de cada interação estabelecida pelos ligantes e suas proteínas. Além disso, essa tabela será utilizada de forma direta para compor as informações que o usuário tem acesso e pode visualizar através da seção *Dataset summary*, disponível na ferramenta web a ser apresentada na Seção 2.14.

Essa tabela é composta pelas seguintes informações: identificador do PDB; cadeia a qual o ligante se liga; o identificador do ligante no PDB; o número do ligante, isto é, o número de sequência do resíduo ou *Residue sequence number* como é definido pelo padrão de formatação de arquivos PDB; agrupamento ao qual o ligante pertence, a frequência de cada tipo de átomo (aceptor, aromático, doador, hidrofóbico, negativo,

positivo e átomos não classificados), o número total de tipos de átomos, a frequência de cada tipo de interação (empilhamento aromático, interação hidrofóbica, interação repulsiva, ponte de hidrogênio e ponte salina) e finalmente o número de interações que o ligante estabelece com a proteína. Tais informações são úteis para a descrição dos tipos de átomos que existem e como eles de fato interagem com a proteína.

Para extrair as informações sobre cada agrupamento, utilizamos os resultados gerados pelo software Ward conforme foi descrito na Seção 2.9. Com relação a cada tipo de átomo presente no ligante, utilizou-se os resultados gerados pelo software Pmapper que possui a capacidade de descrever os tipos de cada átomo do ligante baseado em propriedades químicas tais como pH e carga formal de um átomo tal como foi descrito na Seção 2.7.2. Ressalta-se aqui o fato de que o Pmapper pode fornecer até dois tipos atômicos para um mesmo átomo, assim, um átomo poderia incrementar a contagem de até duas colunas correspondentes aos tipos de átomos. Por exemplo, um átomo definido como aceptor e doador incrementará em 1 o valor da coluna número de átomos aceptores e da coluna número de átomos doadores, além de incrementar em 2 unidades a coluna que armazena o total de tipos de átomos de forma geral.

As interações foram computadas utilizando-se as informações geradas pelos passos descritos na Seção 2.8. Nestes passos, é gerado um arquivo para cada complexo proteína-ligante definido pelo usuário. Nestes arquivos, como já foi descrito, existem as informações sobre cada interação entre o ligante e a proteína e é a partir dessa listagem de interações que são computadas o número de interações para cada tipo, bem como o número total de interações estabelecidas.

2.11 Frequência de tipos de átomos e interações

Nos estudos envolvendo reconhecimento molecular e descoberta de novos fármacos é importante que seja possível identificar e descrever a estrutura atômica do conjunto de ligantes disponíveis em complexo com uma dada proteína e a forma como estes átomos interagem com ela. A estrutura atômica pode revelar possíveis padrões quanto à composição de átomos de um ligante. Buscar padrões na estrutura permite, por exemplo, descrever um farmacóforo, isto é, uma estrutura básica que é comum a todos os ligantes que interagem com uma certa proteína e que corresponde àquela estrutura necessária para que haja o reconhecimento molecular entre a proteína e um ligante.

Da mesma forma, o conjunto de interações estabelecidas permite descrever a forma como se dá o reconhecimento molecular entre uma proteína e seu ligante. Análises estatísticas desse tipo de informação podem ajudar na busca dos padrões

mais frequentes, que são aqueles que mais ocorrem entre todos os ligantes e a proteína; encontrar padrões mais complexos; ou mesmo padrões específicos que são encontrados em apenas algumas interações proteína-ligante.

Portanto, a fim de disponibilizar estas informações aos usuários, são computadas as frequências de cada tipo de átomo e cada tipo de interação no conjunto total de estruturas definidas pelo usuário. Essa informação pode ser calculada utilizando-se os resultados disponíveis na tabela de estatísticas gerais descrita na seção 2.10.

A frequência pode ser descrita como o número total de vezes que se encontrou um dado tipo de átomo ou um dado tipo de interação em um conjunto de dados composto por diversas proteínas e diversos ligantes. Considera-se nessa contagem cada valor possível de átomos ou interações encontradas nesse conjunto de estruturas. Alguns ligantes, por exemplo, podem não ter nenhum átomo do tipo aromático, enquanto que outros ligantes podem ter mais de 1 átomo desse tipo. Assim, a frequência para o tipo de átomo aromático será calculada considerando-se cada um destes valores dispostos na tabela de estatísticas gerais. Por exemplo, se há 4 ligantes que não estabelecem pontes de hidrogênio e 2 ligantes que estabelecem 5 interações desse tipo, a frequência para 0 interações é 4 e para 5 interações será 2.

Da mesma forma, esse cálculo é aplicado para cada tipo de interação a fim de encontrar a frequência de cada um dos tipos possíveis de interações.

2.12 Cálculo da frequência de resíduos interagindo com a proteína

O estudo do reconhecimento molecular envolve não apenas investigar os tipos de átomos e os tipos de interações envolvidas nesse processo, mas também é importante examinar mais detalhadamente quais resíduos estão presentes no sítio de ligação e quais deles são importantes para que haja o reconhecimento de fato. Resíduos importantes para o reconhecimento molecular frequentemente encontram-se interagindo com diversos ligantes, como é o caso dos resíduos Glu81−Leu83 presentes no sítio de ligação da CDK2 e que são importantes para a interação com seus inibidores [Schonbrunn et al., 2013].

Dessa forma, contabilizamos a frequência com que cada resíduo interage com o conjunto de ligantes, a partir de cada interação mapeada em nossa abordagem. Por exemplo, um resíduo que é encontrado interagindo com todos os ligantes terá uma frequência de 100% e um outro que interage com a metade dos ligantes terá uma frequência igual a 50%.

Esse tipo de contagem de frequência funciona bem quando o conjunto de estruturas é composto por uma proteína e vários ligantes, uma vez que sendo a mesma proteína, os resíduos serão os mesmos em todas as estruturas e em todas as posições da sequência. Entretanto, isso não irá ocorrer em todos os casos, visto que a ferramenta permite que o usuário faça pesquisas por estruturas PDBs com diferentes identidades de sequência. Mesmo uma pesquisa com 90% de identidade pode apresentar proteínas em que o sítio de ligação varia em um ou mais resíduos. Dessa forma, a contagem de frequência pode não representar bem a real importância de cada resíduo.

Nesses casos onde a identidade de sequência é inferior a 100%, o que pode ser importante para que as proteínas interajam com o ligante é uma posição em um alinhamento de sequência (na nossa ferramenta, baseado em estrutura) e não um resíduo em específico. Por exemplo, um ligante pode interagir com um conjunto de proteínas porque elas apresentam um aminoácido que contenha anel aromático na posição de uma certa sequência. Assim, dado que as posições da sequência podem conter resíduos diferentes em cada proteína na base de dados, é importante contabilizar a frequência de resíduos de acordo com a posição que estes se encontram em um alinhamento.

Para que isto seja possível, o primeiro passo é realizar um alinhamento de estruturas utilizando o *software* Multiprot[17] [Shatsky et al., 2004]. Uma vantagem desse alinhador é que o resultado gerado por ele apresenta uma correspondência entre as posições da sequência da proteína usada como modelo para o alinhamento e as proteínas alinhadas. Dessa forma, é possível descobrir para uma mesma posição do alinhamento quais são os resíduos encontrados no conjunto todo de estruturas e quais destes resíduos interagem com os ligantes.

Por exemplo, foi feito um alinhamento estrutural com as estruturas contidas nos PDBs 1BGG:D, 1TR1:D, 1E4I:A, 1UYQ:A, 1IFU:A e 3QL8:A, onde a letra após o símbolo ':' representa qual cadeia foi alinhada e sendo a estrutura 1BGG:D o modelo utilizado no alinhamento. Na posição 121 do alinhamento, encontrou-se uma histidina interagindo nos PDBs 1BGG, 1TR1, 1E4I e 1UYQ. Na posição 166, encontrou-se um glutamato interagindo nos PDBs 1BGG, 1E4I e 1UYQ; e uma Valina interagindo no PDB 3QL8. Já a posição 213 apresentou um glutamato interagindo apenas no PDB 1IFU.

Dessa forma, tendo esse mapeamento de posições de sequências concluído, é possível calcular a frequência para cada posição do alinhamento e descobrir quais posições estão envolvidas na interação sem perder, entretanto, as informações detalhadas de quais resíduos foram encontrados nessas posições e quais estruturas possuem tais resí-

[17] http://bioinfo3d.cs.tau.ac.il/MultiProt/

duos.

2.13 Representando interações proteína-ligante através de grafos

Neste trabalho, também disponibilizamos ao usuário uma série de grafos representando as interações proteína-ligante com o objetivo de complementar a análise e auxiliar na descoberta de padrões relevantes que em um primeiro momento sejam difíceis de identificar. Grafos são modelos matemáticos constituídos por um conjunto de objetos chamados vértices (ou nós) e por um conjunto não ordenado de pares de vértices chamado aresta (conexões entre dois vértices). Uma forma simples de entender o que é um grafo é através de uma analogia sobre cidades e rodovias. Nesse contexto, uma modelagem em grafo envolveria a representação de cada cidade como um vértice do grafo e cada rodovia que conecta duas cidades seria uma aresta [Chartrand et al., 2010]. O mapeamento da interface entre uma proteína e seu ligante pode, então, ser gerado de tal forma que vértices representam os átomos da proteína e do ligante que estão em contato (de acordo com suas características químicas e critério de distância); e a as arestas representem as interações formadas por cada par de átomos.

2.13.1 Tipos de grafos

Em teoria de grafos existem diversos tipos de grafos. Aqui serão apresentados alguns dos seus tipos mais comuns.

Grafos podem ser do tipo não-direcionado ou do tipo direcionado (também chamados de dígrafos). Estes termos definem basicamente se existe uma direção ou não nas arestas do grafo [Chartrand et al., 2010].

Grafos podem ainda ser classificados como grafos simples ou multigrafos. Em grafos simples é permitido a existência de apenas uma aresta conectando dois nós, enquanto que em multigrafos podem existir mais de uma aresta conectando dois vértices.

Grafos podem ainda ser classificados como grafos ponderados quando seu conjunto de arestas possuem um peso ("custo") associado a elas [Chartrand et al., 2010].

Um grafo pode ainda ser classificado como grafo rotulado quando seus vértices ou arestas possuem um rótulo associado a eles [Chartrand et al., 2010].

Em uma possível modelagem de proteínas, os vértices podem representar os átomos, os rótulos dos vértices os tipos de cada átomo, as arestas as interações e os rótulos das arestas os tipos das interações entre dois átomos.

Há ainda um tipo particular de grafos chamado bipartido. Nesse tipo de grafo, há dois tipos de nós (conjuntos de nós U e V) e todas as arestas do grafo conectam obrigatoriamente um nó do tipo U a outro do tipo V. Esse é exatamente o tipo de grafo que usamos neste trabalho visto que toda aresta modela uma interação proteína-ligante. Dessa forma, nos nossos grafos o conjunto U poderia ser composto por átomos da proteína e o V, pelos átomos do ligante.

2.13.2 Modelagem para interações proteína-ligante

Os grafos representando as interações proteína-ligante são modelados tal como grafos bipartidos simples, não-direcionados e não ponderados. Além disso, os vértices e as arestas foram rotuladas de acordo com o tipo do átomo e o tipo da interação, sendo que alguns átomos podem ser classificados em mais de um tipo. Por exemplo, o átomo OE2 do aminoácido Glutamato é rotulado como aceptor e negativo (Tabela 2.2). Já as arestas, recebem sempre um único rótulo referente à interação estabelecida de acordo com a ordem de prioridade definida para as interações (Seção 2.8).

A representação das interações em formato de grafo foi feita utilizando-se o formato Graphml que é um formato para representação de grafos baseado no XML. Esse formato é útil, pois permite associar diversos atributos aos nós e arestas, de acordo com a modelagem aqui proposta para grafos proteína-ligante. Em nossa modelagem utilizamos como atributos específicos os seguintes itens: o tipo de cada átomo e interação, o número de série do átomo (*Atom serial number*) no PDB e as cores para cada nó de acordo com a molécula que o contém – nós azuis pertencem à proteína e nós laranjas pertencem ao ligante. As arestas são rotuladas de acordo com os seguintes tipos de interação: AR (Aromatic stacking), HB (Hydrogen bond), HP (Hydrophobic interaction), RP (Repulsive interaction) e SB (Salt bridge).

Imagens de tais grafos são geradas a partir dos arquivos Graphml e ficam disponíveis para que o usuário possa visualizá-los e analisá-los na ferramenta *web*. Esse procedimento é feito utilizando-se a ferramenta Graph-tool[18], que é um módulo Python disponível para manipulação e análise estatística de grafos. Todas as imagens estão disponíveis para acesso a partir da seção *Protein-ligand interactions by ligands* do nAPOLI. Um exemplo de grafo construído com essa abordagem é ilustrado na Figura 2.8 a seguir.

Assim, por meio destas informações o usuário poderá: analisar todos os complexos em busca de padrões relevantes, constatar quantos e quais átomos estão envolvidos na interação e descobrir a topologia do grafo, revelando, por exemplo, conexões que pos-

[18]https://graph-tool.skewed.de/

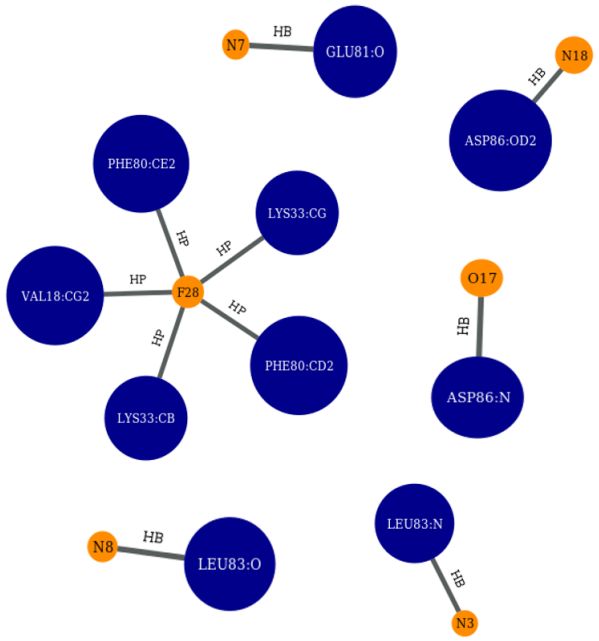


Figura 2.8. Exemplo de grafo para o PDB 3R9H. Nós azuis e laranjas representam os átomos da proteína e do ligante, respectivamente. As interações entre dois átomos são representadas por arestas.

sam ser importantes para o reconhecimento molecular entre as moléculas envolvidas. Em trabalhos futuros, essa modelagem será utilizada para a computação de padrões frequentes de forma automática conjuntamente com uma abordagem de mineração em grafos para revelar padrões não óbvios, facilitando, portanto, o estudo e descoberta de quais átomos e interações são as mais frequentes e importantes. Consequentemente, essas informações podem revelar átomos e interações necessárias para que haja o reconhecimento entre uma proteína e seus ligantes ou vice versa.

Entretanto, um problema enfrentado com a representação dos grafos surgiu devido às interações hidrofóbicas e empilhamentos aromáticos. Nossa ferramenta possibilita que a escolha dos parâmetros para o cálculo de interações seja feita de forma customizada pelo usuário. Nesse ponto, à medida que se aumentam os limiares de distância para uma interação, aumenta-se também a chance de encontrar interações de um dado tipo com átomos mais distantes. Dessa forma, o número de interações que podem ser estabelecidas quando se aumentam as distâncias para interações hidrofóbicas ou empilhamentos aromáticos aumenta de tal maneira que impossibilita a análise visual do grafo, como pode ser visto na Figura 2.9(a). Com relação às interações hidrofóbicas, isso ocorre uma vez que existe uma grande quantidade de átomos hidrofóbicos em uma proteína (Tabela 2.2), permitindo que diversas interações possam ser consideradas à medida que se aumenta a distância. Já no caso de empilhamentos aromáticos, isso

ocorre porque os anéis aromáticos dos resíduos fenilalanina (6 átomos aromáticos), histidina (5 átomos aromáticos), tirosina (6 átomos aromáticos) e triptofano (9 átomos aromáticos) apresentam muitas opções de interações desse tipo em um único resíduo.

Outro problema que surgirá nos nossos trabalhos futuros de mineração de padrões em grafos será o aumento no número de interações estabelecidas que provoca um enorme crescimento no número de subgrafos possíveis (subconjunto de vértices e arestas de um grafo). Isso permite que sejam encontrados muitos grafos isomorfos, o que aumenta a complexidade do problema. Dois grafos G e H são ditos isomorfos se eles possuem a mesma estrutura, de forma que todos os vértices adjacentes (conectados) em G são vértices adjacentes em H. Como o escopo desse trabalho não foi a mineração de padrões, não entraremos em detalhes sobre esse problema e sua complexidade. Para mais referências, ver Chartrand et al. [2010]; Zaki e Meira [2014].

Tendo essas dificuldades em mente e em especial a dificuldade de se visualizar grafos densos, uma forma de tratar ambos os problemas citados é com a sumarização de conjuntos de átomos com mesma característica a partir do cálculo de centroides. Esse procedimento foi conduzido para as interações hidrofóbicas e empilhamentos aromáticos. Utilizou-se uma abordagem similar àquelas utilizadas em cálculos de contatos que empregam a ideia de centroides para simplificar o problema a nível de resíduo. Geralmente, as abordagens para cálculo de contatos utilizam algum dos seguintes representantes de um resíduo: carbonos-α (CA), carbonos-β (CB), centro geométrico (GC) ou baricentro (BC). Porém, nossa metodologia não leva em consideração a posição de um centroide, ao contrário destes métodos citados anteriormente em que a posição faz toda a diferença. Isto ocorre, pois a modelagem de grafos que utilizamos leva em consideração apenas as informações dos átomos (nós) e as interações (arestas) que existem entre dois pares de átomos, ou seja, a distância e o posicionamento dos nós no grafo não refletem as distâncias e posições reais dos átomos na estrutura.

A ideia por trás dessa abordagem é a redução do número de arestas e nós, simplificando, portanto, o grafo como um todo. Isso permite que as imagens geradas fiquem mais legíveis e que haja uma redução do número de subgrafos. Para isso, primeiramente, verifica-se se existe pelo menos um resíduo que faça mais de uma interação hidrofóbica ou empilhamento aromático. Se sim, será criado um novo nó (centroide) para representar o resíduo. Em seguida, todas as interações hidrofóbicas ou empilhamentos aromáticos serão remapeados sobre o centroide. Apesar de um único centroide ser necessário para representar o resíduo, as interações são mapeadas de forma isolada; isto é, se existirem resíduos que façam mais de uma interação hidrofóbica e não existirem resíduos que façam mais de uma interação do tipo empilhamento aromático, o mapeamento será feito apenas para as interações hidrofóbicas. Além disso, os cen-

troides criados são rotulados de acordo com o nome e número do resíduo no PDB e o tipo do centroide. O tipo é uma sigla que indica que o nó representa um centroide e qual resíduo ele representa. Isto é feito seguindo o padrão *COR* que significa *Centroid of Residue*, em que *Residue* deve ser substituído pelo código de uma letra do resíduo. Por exemplo, o centroide de uma histina será rotulado como COH (*Centroid of Histidine*); o de uma leucina será COL; e o de uma tirosina será COY e assim por diante.

A Figura 2.10 ilustra esse processo. Note que em 2.10(a), há dois grupos de resíduos: o grupo destacado pelo círculo em ciano apresenta três interações hidrofóbicas do resíduo tirosina 435 com dois átomos do ligante (C e CB); já o grupo em destaque pelo círculo vermelho apresenta cinco interações do resíduo fenilalanina 438 com dois átomos do ligante (CB e CG). Sendo todas essas interações do tipo interação hidrofóbica (HP). Conforme explicado anteriormente, ambos os resíduos estabelecem mais de uma interação hidrofóbica com o ligante e, portanto, todas essas interações devem ser remapeadas. Em 2.10(b), é possível ver os centroides criados para cada resíduo. Note que o centroide (círculo em ciano) para a tirosina (COY) manteve as interações com cada átomo do ligante, sendo que as interações que eram estabelecidas com o átomo CB agora é representada por uma única interação. O mesmo ocorre para o centroide da fenilalanina (COF), porém, esse resíduo estabelecia duas interações com o átomo CB e 3 interações com o átomo CG, de modo que após o mapeamento restou apenas uma interação para cada um desses átomos. Vale ressaltar também que em ambos os grafos, as demais interações não sofreram modificações.

Esse exemplo representa um caso didático apenas, pois o grafo sem centroide não apresenta nenhuma ilegibilidade. O exemplo foi dado apenas para descrever o funcionamento do algoritmo. Porém, existem casos mais complexos em que os centroides são necessários. Veja e compare os grafos na Figura 2.9. Note que a adição de centroide (2.9(b)) permite a análise mais precisa de quais resíduos da proteína e quais átomos do ligante estão interagindo.

2.14 A ferramenta web

Toda as opções disponíveis para pesquisa e os resultados das análises de padrões proteína-ligante podem ser acessadas pelo usuário através da página web do nAPOLI[19]. Essa página possui duas seções principais: a busca de PDBs (*PDB search*) e o acesso aos resultados das análises de padrões proteína-ligante (*Results*).

[19]http://www.napoli.dcc.ufmg.br/

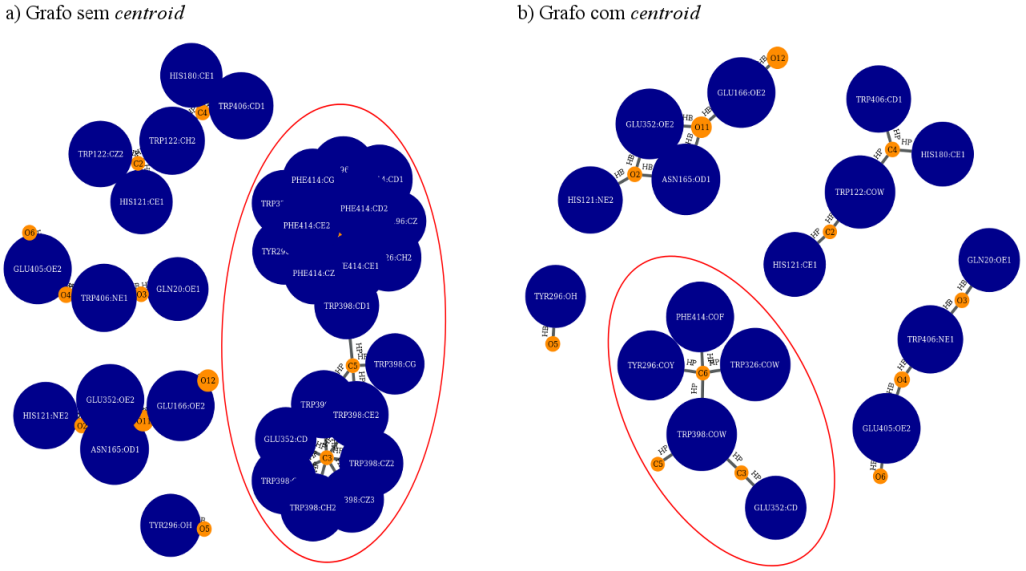


Figura 2.9. Cálculo de centroides para o PDB 1BGG em complexo com o ligante Ácido glucónico (GCO). Os círculos em vermelho representam o mesmo subgrafo. Em (a) não foi utilizado o algoritmo de mapeamento por centroides e com isso o subgrafo em destaque ficou ilegível para análises. Já em (b), a partir do uso desse algoritmo é possível ver que o mesmo subgrafo se tornou legível.

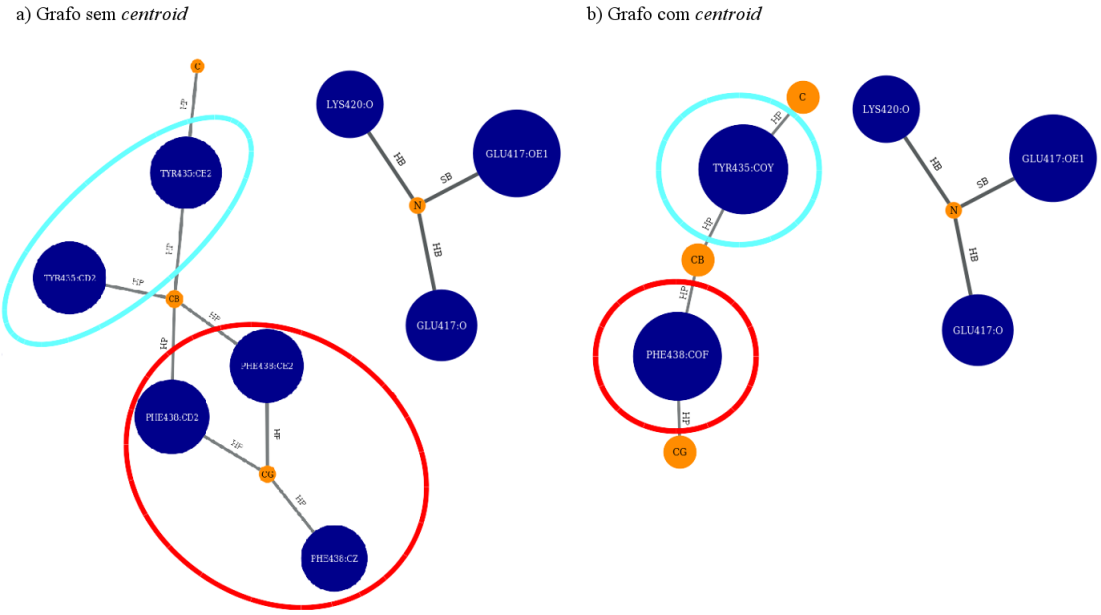


Figura 2.10. Cálculo de centroides para o PDB 1XDJ em complexo com o ligante Selenometionina (MSE). Em (a) o grafo foi representado sem o uso de centroides e em (b) foram gerados os centroides para o resíduo tirosina 435 (círculo em ciano) e para o resíduo fenilalanina 438 (círculo vermelho).

Na seção de pesquisa de PDBs, o usuário tem a opção de buscar por estruturas de proteínas e grupos de ligantes que estejam em complexo com as proteínas encontradas. Dessa forma, o usuário pode compor sua base de dados consistindo de diversas proteínas e ligantes (conforme descrito na Seção 2.4). Tendo escolhido sua base de dados, o usuário poderá submeter essas informações para serem processadas em nosso servidor.

Todos os resultados obtidos com a análise das bases de dados do usuário encontram-se disponíveis em Resultados (*Results*). É nessa seção que ele poderá conferir, interpretar e estudar os padrões proteína-ligante presentes na base de dados submetida.

Além destas seções, o usuário também tem acesso a um manual (*Help*) que tem como objetivo auxiliá-lo a desfrutar e ter a melhor experiência possível com o nAPOLI.

Todas as páginas foram construídas através da linguagem PHP e CodeIgniter[20] que é um *framework* para PHP que auxilia e simplifica a construção de páginas PHP. Além disso, em todas as páginas utilizou-se *scripts* JavaScript que é uma linguagem de programação que permite construir páginas interativas. A principal biblioteca JavaScript utilizada é o jQuery[21] que fornece diversas funcionalidades simplificadas para programação em JavaScript e a biblioteca D3.js[22] [Bostock et al., 2011] que oferece uma ampla gama de aplicações para visualização de dados.

2.15 Estratégias visuais para a análise de padrões proteína-ligante

Nesta sessão, serão descritas as estratégias visuais adotadas para auxiliar no estudo dos padrões proteína-ligante. Cada uma das estratégias foram propostas para responder às seguintes questões:

1. Quais são as possíveis interações que cada ligante pode estabelecer com a proteína e como de fato isso ocorre?
2. Qual é a frequência de cada tipo de átomo no conjunto de ligantes?
3. Qual é a frequência de cada tipo de interação no conjunto de ligantes?
4. Quais são os resíduos que frequentemente interagem com os ligantes?
5. Existem grupos de ligantes similares?

[20] http://www.codeigniter.com/
[21] https://jquery.com/
[22] http://d3js.org/

Além disso, para cada estratégia visual, o usuário terá a opção de realizar as análises para o conjunto completo de dados ou realizar a análise de acordo com cada agrupamento de ligantes similares.

2.15.1 Resumo do conjunto de dados (*Dataset summary*)

Nesta página, disponibilizamos uma tabela interativa que contém um resumo sobre o conjunto de dados. Nela, estão presentes as seguintes informações: o identificador PDB onde foi encontrado o ligante, a cadeia na qual o ligante se encontra, o identificador do ligante, o número que o ligante recebeu no arquivo PDB, o agrupamento (*cluster*) ao qual o ligante pertence, o número de átomos do ligante e o número de interações estabelecidas pelo ligante 2.11.

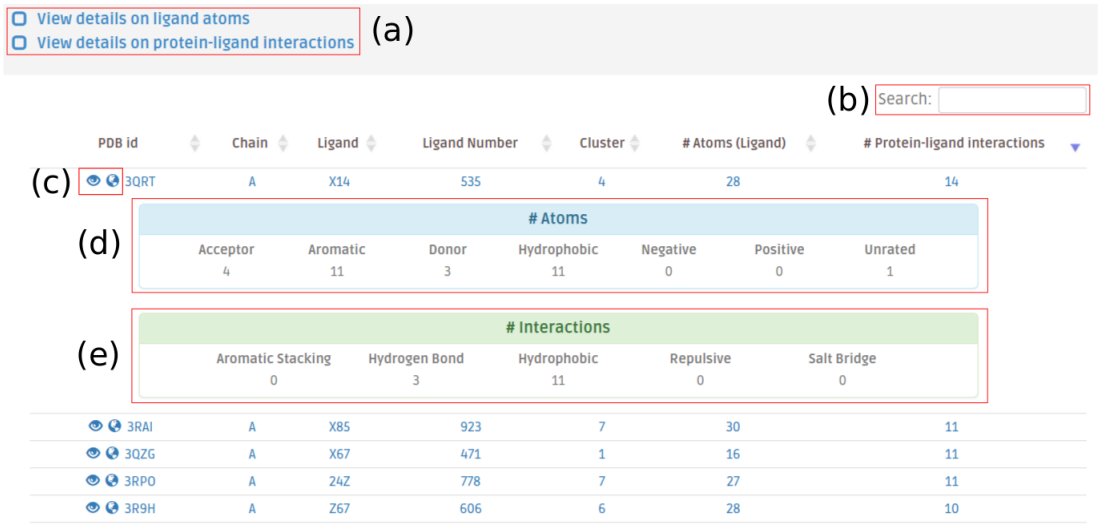


Figura 2.11. Tabela disponível em *Dataset summary* (*Resumo do conjunto de dados*). (a) Campo de seleção (*checkbox*) para mostrar automaticamente todos os painéis de tipos de átomos ou interações. (b) Campo de pesquisa para filtrar informações. (c) Botões para visualizar a estrutura tridimensional do complexo de forma interativa ou acessar a página referente à uma estrutura diretamente no site PDB. (d) e (e) Exemplos de painéis de tipos de átomos e interações, respectivamente.

Para proporcionar a interatividade com o usuário, estão disponíveis diversas opções de filtragem, ordenação e detalhes sob demanda.

Com relação às opções de filtro, está disponível um campo de pesquisa no qual o usuário pode selecionar informações específicas que ele esteja interessado em um certo momento 2.11(b). Por exemplo, alguém poderia estar interessado em pesquisar ligantes que tenham 16 átomos, então, bastaria digitar o valor 16 nesse campo que au-

tomaticamente todas as linhas que contivessem esse valor seriam mostradas. Por outro lado, se ele estiver interessado em analisar um único PDB bastaria que ele digitasse o identificador desse PDB no campo de pesquisa e somente as linhas que tivessem esse identificador seriam mostradas na tabela.

Além disso, estão disponíveis diversas opções de ordenação dos dados, seja de forma ascendente ou descendente. Para isso o usuário deverá apenas clicar sobre o cabeçalho da coluna que ele deseja ordenar. Por exemplo, alguém poderia estar interessado em ordenar os dados em ordem crescente de acordo com o número de átomo ou de interações estabelecidas por um ligante.

Também estão disponíveis diversas opções para obter detalhes sob demanda. Por exemplo, clicando sobre um valor da coluna de *clusters* será mostrada uma imagem da sobreposição de todos os ligantes desse grupo. A sobreposição foi feita através de um alinhamento múltiplo de estruturas utilizando o *software* LovoAlign [Martínez et al., 2007] e a imagem da sobreposição foi gerada automaticamente através do *software* PyMOL [Schrödinger, LLC, 2010]. LovoAlign é útil nesse processo porque cria novos arquivos PDB em que as coordenadas dos átomos de todas as estruturas alinhadas são alteradas para equivaler às coordenadas da estrutura modelo. Assim, com o alinhamento das estruturas todos os ligantes de uma região da proteína estarão sobrepostos. Os ligantes são mostrados em bastões (*sticks*) e são coloridos de acordo com o esquema de cores CPK. Essas imagens são úteis, pois permite ao usuário visualizar e constatar quais são as as estruturas químicas similares em cada agrupamento 2.12(a).

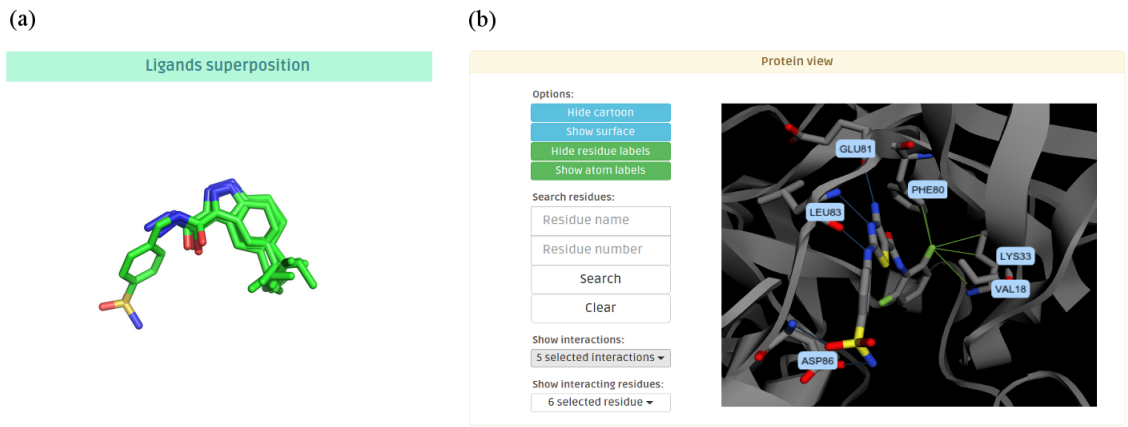


Figura 2.12. Visualizações disponíveis na tabela *Dataset summary* (*Resumo do conjunto de dados*). (a) Sobreposição dos ligantes pertencentes a um grupo de moléculas similares. (b) Visualização da estrutura tridimensional contendo diversas opções de interação com o sítio de ligação.

Clicando sobre os valores presentes na coluna *# Atoms (ligand)* (*Número de*

átomos), surge um painel contendo informações úteis e detalhadas sobre cada tipo de átomo presente no ligante. Nesse painel, é possível analisar quantos átomos aceptores, aromáticos, doadores, hidrofóbicos, negativos, positivos e quantos átomos não foram classificados em nenhum desses tipos pelo Pmapper 2.11(d).

Da mesma forma, se o usuário clicar sobre algum valor da coluna *# Protein-ligand interactions* (*Número de interações proteína-ligante*) será mostrado um painel contendo informações sobre cada tipo de interação estabelecida entre a proteína e o ligante. As interações disponíveis são: empilhamento aromático, ponte de hidrogênio, interação hidrofóbica, interação repulsiva e ponte salina 2.11(e).

Para facilitar a análise das informações sobre o número de átomos e o número de interações proteína-ligante, existe um campo (*checkbox*) que o usuário pode selecionar e mostrar de uma única vez todos os painéis referentes aos tipos de átomos existentes ou tipos de interações estabelecidas (Figura 2.11(a)).

Além disso, para complementar as análises o usuário tem disponível ainda uma visualização na qual ele pode interagir diretamente com o sítio de ligação de forma similar àquela disponível no PyMOL (Figura 2.12(b)). Nessa visualização o usuário poderá rotacionar a câmera e analisar a estrutura em diferentes ângulos; utilizar o zoom para dar ênfase à alguma parte específica da estrutura; mostrar ou esconder a representação em cartoon da estrutura da proteína; mostrar ou esconder a superfície da proteína de acordo com os raios de van der Waals; analisar em detalhes quais tipos de interações estão ocorrendo (linhas conectando dois átomos); mostrar ou esconder os rótulos de átomos e resíduos; pesquisar, mostrar ou esconder resíduos específicos, entre outras coisas. Essa visualização foi construída utilizando a biblioteca JavaScript 3Dmol.js[23] [Rego e Koes, 2015]. Para acessar essa visualização basta que o usuário clique com o mouse sobre o ícone em forma de "olho" disponível em cada linha dessa tabela, então, um painel contendo a visualização será aberto (Figura 2.11(c)).

O usuário poderá ainda acessar maiores informações sobre uma proteína clicando sobre o ícone em forma de "globo terrestre". Esse ícone abre, em uma nova aba no navegador, as informações referentes a uma estrutura diretamente através do site do PDB[24] (Figura 2.11(c)).

Finalmente, o usuário pode optar tanto por analisar todo o conjunto de dados ou analisar apenas os dados referentes a um agrupamento de ligantes similares (*clusters*).

[23] http://3dmol.csb.pitt.edu/index.html
[24] http://www.rcsb.org/pdb/home/home.do

2.15.2 Análises gráficas (*Graphical analysis*)

Em *Graphical analysis* os usuários poderão realizar suas análises através de diversos tipos de gráficos e formas variadas de interação com as informações disponíveis. São aplicadas diversas técnicas de visualização de dados para maximizar a experiência do usuário com a ferramenta. Nela o usuário poderá ordenar os valores, selecionar informações específicas, analisar o conjunto todo de informações, escolher o tipo de gráfico desejado e obter detalhes sob demanda. As análises foram separadas basicamente em três classes: tipos de átomos, tipos de interações e correlação entre o número de átomos e número de interações estabelecidas para um dado tipo.

Os gráficos dessa seção foram construídos utilizando a biblioteca JavaScript D3.js[25] [Bostock et al., 2011].

2.15.2.1 Tipos de átomos

Nessa primeira classe de páginas, o usuário poderá analisar as informações referentes às frequências dos tipos de átomos no conjunto completo de ligantes ou analisar as frequências destes átomos de acordo com cada agrupamento (Figura 2.13(a)). Essas páginas podem ser acessadas respectivamente em *Atoms type (all data)* (*Tipos de átomos (todos os dados)*) e *Atoms type (clusters)* (*Tipos de átomos (grupos)*). Em ambas as páginas os usuários terão acesso às distribuições de tipos de átomos de um dado tipo no conjunto todo de ligantes ou de acordo com o grupo ao qual ele pertence.

Estão disponíveis três tipos de gráficos: gráficos de barras, gráficos de pizza e gráficos de barras agrupadas (apenas na seção de *clusters*).

Através dos gráficos de barras (Figura 2.13), os usuários poderão analisar a forma como um conjunto de valores são distribuídos ao longo de um intervalo e descobrir possíveis padrões visuais como o espalhamento, a centralidade e a forma dos dados, isto é, se a forma é assimétrica ou simétrica, se há múltiplos picos ou não, se a forma é curva ou plana. No eixo X são mostrados diferentes números de átomos para um dado tipo e no eixo Y mostra-se quantos ligantes possuem aquela quantidade de átomos (frequência). Junto ao gráfico de barras também é adicionado um gráfico de Pareto que é um tipo de gráfico de linha que mostra a frequência acumulada dos dados em porcentagem. O gráfico de Pareto é útil, pois revela os valores mais importantes, que são aqueles valores com grande número de representantes. Em nosso caso, os valores mais importantes seriam os números de átomos de um dado tipo que são mais frequentemente encontrados ao longo do conjunto de ligantes. Além disso, os usuários podem interagir com os gráficos de forma a obter informações detalhadas (valores

[25] http://d3js.org/

precisos) quando estes passam o mouse sobre cada barra ou cada ponto do gráfico de linha (Figura 2.13(d)). O usuário também tem a opção de analisar todos os tipos de átomos de uma só vez, nesse caso os gráficos de barras são posicionados um ao lado do outro em um tamanho reduzido para permitir que todos eles estejam visíveis na tela.

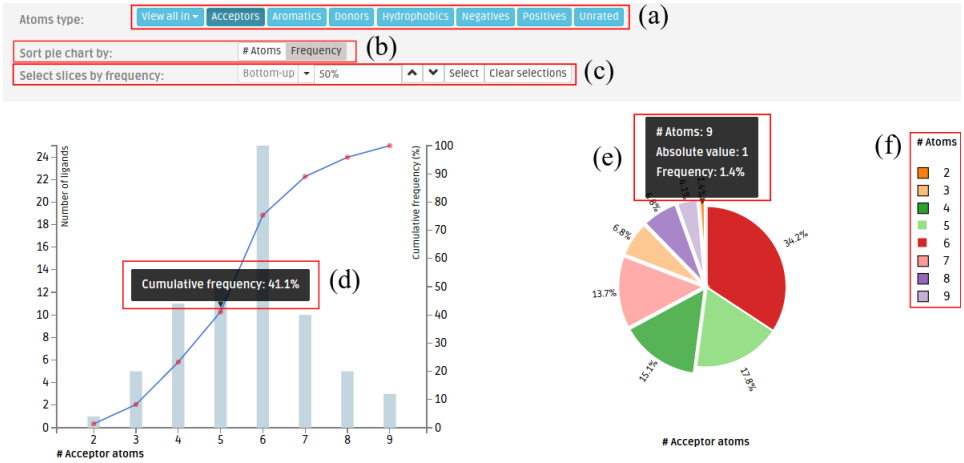


Figura 2.13. Gráficos de barras e pizza disponíveis na seção de *Graphical analysis*. Em (a) é mostrado as opções de tipos de átomos que podem ser selecionadas. Neste exemplo, o tipo de átomo Aceptor foi selecionado. (b) Opções de ordenação das fatias do gráfico de pizza. (c) Opções de seleção das fatias de acordo com a frequência. (d) e (e) Exemplo de interatividade (valores sob demanda) com o gráfico de barras e pizza, respectivamente. (f) Legenda interativa do gráfico de pizza.

Os gráficos de pizza (Figura 2.13) complementam as análises e mostram o relacionamento da parte e o todo, isto é, cada fatia do gráfico de pizza representa a proporção de um valor em relação ao todo. Assim, fatias maiores correspondem àqueles números de átomos que possuem uma alta frequência. Porém, segundo Few [2009], comparar as fatias dos gráficos de pizza é algo difícil já que não é possível mensurar com exatidão as áreas e os ângulos formados por cada fatia. Apesar disso, esses gráficos são bons complementos para as análises. Da mesma forma, o usuário pode interagir com estes gráficos de diferentes maneiras. Ele poderá obter valores sob demanda tal como ocorre para os gráficos de barras (Figura 2.13(e)). Ao clicar sobre uma fatia, seu rótulo será escondido ou mostrado de acordo com a necessidade que pode ser destacar apenas os valores desejados ou facilitar a visualização quando os rótulos se sobrepõem. O usuário também pode ordenar as fatias de acordo com o número de átomos ou de acordo com a frequência (número de ligantes que apresentam uma determinada quantidade de átomos) (Figura 2.13(b)). O usuário também pode definir um valor em porcentagem que será utilizado para selecionar as fatias cuja frequência

quando somadas não ultrapassem o valor definido (Figura Figura 2.13(c)). As fatias selecionadas serão destacadas para facilitar a visualização. Isso é útil, por exemplo, para encontrar as fatias que representam 50% dos dados. Essa busca pode ser feita tanto das fatias de maior frequência para as de menor frequência (*top-down*) quanto da menor frequência para a maior (bottom-up). O usuário poderá ainda destacar ou remover o destaque das fatias clicando sobre as legenda do gráfico de pizza (Figura 2.13(f)). Além disso, também é possível visualizar todos os tipos de átomos de uma única vez tal como descrito para os gráficos de barra.

Os gráficos de barras agrupadas (Figura 2.14), por outro lado, estão presentes apenas nas páginas de análise de padrões em grupos de ligantes similares. Esses gráficos permitem que os usuários comparem a distribuição do número de átomos em cada agrupamento de ligantes, sendo possível, portanto, analisar se existe um padrão referente ao número átomos para cada grupo de ligantes. De forma similar, os usuários podem obter valores sob demanda à medida que passam o mouse sobre uma barra.

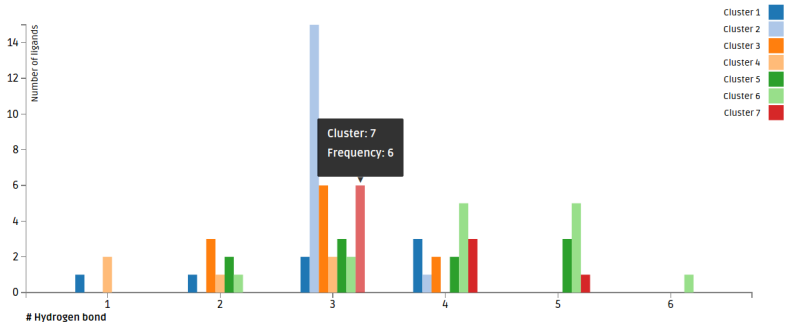


Figura 2.14. Exemplo de gráfico de barras agrupadas para a análise de grupos (*clusters*) de ligantes. Cada barra representa um grupo diferente.

2.15.2.2 Tipos de interações

A segunda classe de páginas permite ao usuário analisar as informações referentes às frequências dos tipos de interações no conjunto completo de ligantes ou analisar as frequências destas interações de acordo com cada agrupamento. Essas páginas podem ser acessadas respectivamente em *Tipos de interações (todos os dados)* (*Interactions type (all data)*) e *Tipos de interações (grupos)* (*Interactions type (clusters)*). Em ambas as páginas os usuários terão acesso às distribuições de tipos de interações de um dado tipo no conjunto completo de ligantes ou de acordo com o grupo ao qual ele pertence. Todas os gráficos e opções disponíveis para número de átomos também estão disponíveis nessa classe de páginas.

2.15.2.3 Correlação entre número de átomos e número de interações

A terceira classe, diferentemente das outras duas, tem como objetivo auxiliar na comparação e investigação de possíveis correlações entre o número de átomos e o número de interações de um dado tipo de forma visual (Figura 2.15). É possível, por exemplo, verificar se à medida que se aumenta o número de átomos hidrofóbicos também se verifica um aumento no número de interações hidrofóbicas. Para tanto, disponibilizamos uma série de gráficos de pontos (*scatterplot*) nos quais no eixo X estão os números de átomos de um dado tipo e no eixo Y o número de interações encontradas para um tipo de átomo. No caso dos átomos carregados que possuem dois tipos de interações possíveis, mostra-se ambos os tipos interações em um mesmo gráfico, de forma que o símbolo utilizado para representar os pontos de interações repulsivas é um quadrado e para as pontes salinas é um círculo. De forma similar, os usuários também podem obter valores sob demanda. Esses gráficos encontram-se disponíveis em *Atoms x Interactions*.

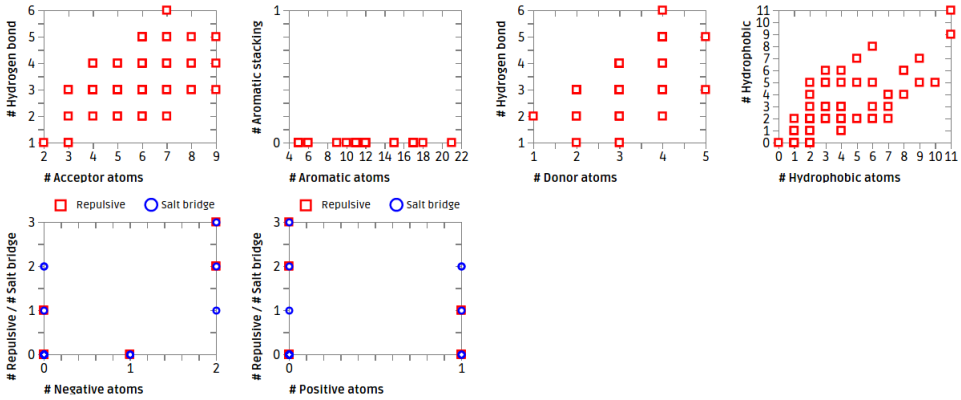


Figura 2.15. Exemplo de gráficos de correlação *Atoms x Interactions*. Nesta figura, são mostrados todos os gráficos de correlação, ou seja, um gráfico para cada tipo de átomo e as interações possíveis para estes átomos. Essa visualização pode ser selecionada a partir do botão *View all* (*Ver todos*).

2.15.3 Interações por resíduos (*Interactions by residues*)

No estudo do reconhecimento molecular é importante examinar mais detalhadamente quais resíduos estão presentes no sítio de ligação, quais resíduos mais frequentemente encontram-se interagindo com o conjunto todo de ligantes e quais os tipos de interações que frequentemente são estabelecidas por esses resíduos porque tais informações permitem-nos desvendar e entender os elementos essenciais para que o reconhecimento ocorra de fato.

Para auxiliar nessa descoberta, foi proposta uma tabela codificada por cores e interativa (Figura 2.16). Essa tabela possui três colunas principais: nome do átomo (*Atom*), tipo de interação (*Type of interaction*) e número de ligantes com o qual um resíduo interage (*# Ligands with which it interacts*). Clicando em cada uma destas colunas o usuário pode ordenar os dados de acordo com a informação desejada.

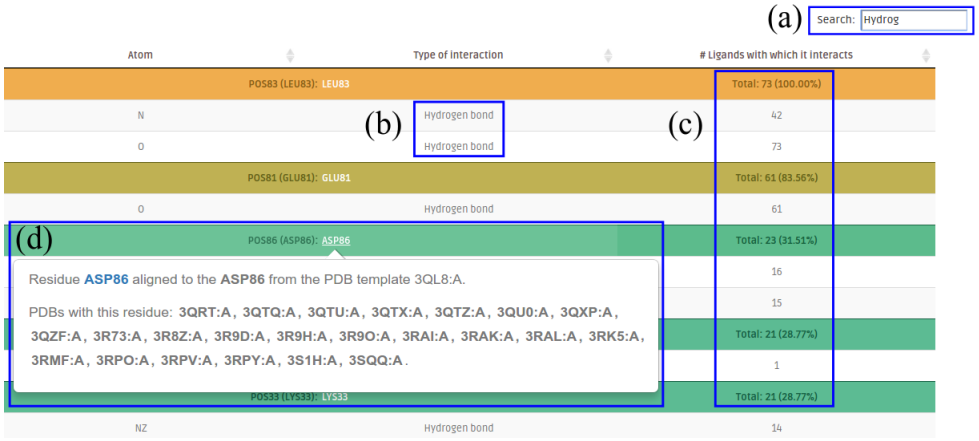


Figura 2.16. Tabela codificada por cores da seção *Interactions by residues*. (a) Campo de pesquisa para filtrar informações na tabela. (b) Resultado do filtro aplicado após ser digitado "Hydrog" no campo de seleção. (c) Células de Frequência total e valores de frequência para diferentes resíduos. (d) Células de Posições do alinhamento ou nome do resíduo e caixa de texto mostrando as informações do alinhamento e estruturas que possuem o resíduo interagindo com algum ligante.

Além dessas colunas, ainda existem as informações das posições do alinhamento (rever Seção 2.12) ou nome do resíduo (Figura 2.16(d)) – o nome do resíduo somente será utilizado quando uma posição da sequência de aminoácidos de uma proteína não puder ser alinhada com a estrutura modelo utilizada no alinhamento – e a frequência total (Figura 2.16(c)) na qual uma posição do alinhamento ou resíduo foi encontrada interagindo com o conjunto todo de ligantes. Essas informações são utilizadas para agrupar as linhas das três colunas citadas anteriormente e são dispostas de tal maneira que elas também se comportam como cabeçalhos da tabela. Com isso, é possível verificar para uma determinada posição do alinhamento (ou resíduo) quais são os átomos e os tipos de interações estabelecidas por cada átomo. Além disso, esses cabeçalhos utilizados para agrupar as linhas são codificadas por um sistema de cores que varia do azul (frequências menores) ao laranja (frequências maiores) facilitando, então, a identificação das posições mais frequentes de forma fácil e interativa. É possível ordenar a tabela de acordo com o número da posição do alinhamento ou número do resíduo (quando disponível), para isso basta que o usuário clique sobre a célula dessas infor-

mações. Por outro lado, ao clicar sobre a célula de frequência total ordenará de forma crescente ou decrescente segundo o valor da frequência. Na Figura 2.16(c), as linhas foram ordenadas em ordem decrescente de acordo com a frequência total. A ordenação por frequência é muito útil principalmente para descobrir quais são os resíduos e posições que mais frequentemente interagem com os ligantes.

Conforme explicado na Seção 2.12, uma mesma posição do alinhamento pode apresentar diferentes resíduos, uma vez que nem sempre as proteínas serão 100% idênticas. Assim, nessa tabela é possível verificar a frequência para cada posição do alinhamento e quais são os tipos de interações que são comumente encontradas entre essas posições e os ligantes. Mas, mais que isso, o usuário poderá constatar quais foram os resíduos alinhados em uma mesma posição e quais estruturas possuem esses resíduos. Passando o mouse sobre cada resíduo será aberto uma caixa de texto contendo as informações do alinhamento tais como qual foi a estrutura utilizada como modelo e quais são as estruturas que possuem o resíduo em foco (Figura 2.16(d)).

Outra opção disponível é a filtragem de informações através de um campo de pesquisa disponível no topo da tabela. O usuário poderá digitar um conjunto de caracteres e somente as linhas que contenham essa informação serão mostradas de forma imediata. Por exemplo, se o usuário digitar "Hydrog" (Figura 2.16(a)) somente as linhas que contenham "Hydrogen Bond" estarão visíveis (Figura 2.16(b)). Por outro lado, se ele digita "CB", apenas linhas contendo carbonos β estarão visíveis.

O usuário pode ainda analisar a frequência de resíduos interagindo com os ligantes para cada agrupamento em específico ao invés de verificar o conjunto completo de ligantes. Isso é interessante porque nem sempre um padrão global pode emergir, sendo necessário, então, analisar as interações que grupos de ligantes similares realizam com as proteínas.

Essa tabela foi construída com o auxílio da biblioteca JavaScript DataTables[26] que é uma biblioteca que estende as funcionalidades da biblioteca jQuery.

2.15.4 Interações por ligantes (*Interactions by ligands*)

Da mesma forma que em *Interactions by residues*, propomos uma tabela nesta seção que também tem como objetivo auxiliar na descoberta de quais são os resíduos, quais átomos e quais interações são estabelecidas no conjunto de dados (Figura 2.17). Porém, o foco dessa tabela é apresentar uma visão detalhada sobre cada complexo proteína-ligante, ao contrário da abordagem proposta em *Interactions by residues* em que o foco foi apresentar estatísticas sobre os resíduos mais frequentes.

[26] https://datatables.net/

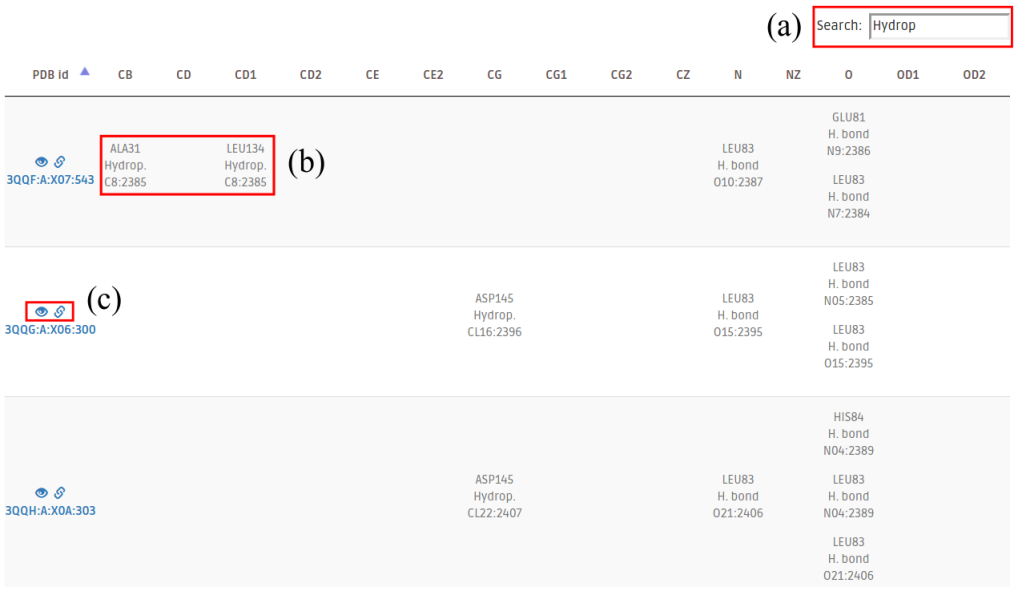


Figura 2.17. Tabela disponível em *Interactions by ligands* que tem como foco descrever as interações estabelecidas por cada ligante. (a) Campo de pesquisa para filtrar informações na tabela. (b) Células com informações sobre as interações estabelecidas com os átomos CB e CD1. (c) Botões para visualizar a estrutura tridimensional do complexo de forma interativa ou os grafos representando as interações proteína-ligante.

Essa tabela é composta basicamente de uma coluna para o *PDB id* (*identificador PDB*) e uma coluna com o nome de cada átomo dos resíduos que interagem com os ligantes. Dessa forma, cada base de dados terá seu conjunto de colunas de átomos específicas, tendo em vista que os átomos que interagem em diferentes complexos serão diferentes. Cada linha irá apresentar um complexo de acordo com o identificador PDB e todas as interações estabelecidas nesse complexo. Para isso, em cada célula (coluna de átomos) terá informado o nome do resíduo que interage com o ligante, o tipo de interação e qual o átomo do ligante que interage com a proteína (Figura 2.17(b)). Se não houver interação entre um ligante e um determinado átomo do resíduo, a coluna referente a esse átomo estará vazia.

Além disso, através dessa tabela, os usuários terão acesso a uma visualização na qual eles podem interagir diretamente com o sítio de ligação de forma similar àquela disponível no PyMOL [Schrödinger, LLC, 2010] (Figura 2.12(b)). A diversidade de interações possíveis com essa visualização foi descrita na Seção 2.15.1. Para acessar essa visualização basta que o usuário clique com o mouse sobre o ícone em forma de "olho" disponível em cada linha dessa tabela, então, um painel contendo a visualização será aberto (Figura 2.17(c)). Da mesma forma, os usuários também poderão acessar os grafos representando as interações proteína-ligante para cada complexo (Figura 2.8).

Nesses grafos, os átomos da proteína são representados por nós azuis, os átomos do ligante são representados por nós laranjas e as interações são as arestas do grafo. Essa visualização pode ser acessada através de um ícone simbolizando um grafo, que está localizado ao lado do ícone descrito anteriormente.

Nessa tabela, o usuário poderá ainda filtrar as informações utilizando um campo de pesquisa no topo da tabela. Esse filtro permite que usuário foque apenas nas informações que ele deseja analisar em um dado momento. Ele poderá digitar "LEU" e somente os ligantes que interagem com uma Leucina serão mostrados. Se ele digitar "Hydrop", o foco será apenas sobre os ligantes que interajam através de interações Hidrofóbicas (Figura 2.17(b)).

Essa tabela foi construída com o auxílio da biblioteca JavaScript DataTables[27] que é uma biblioteca que estende as funcionalidades da biblioteca jQuery.

[27]https://datatables.net/

Capítulo 3

Resultados

O principal resultado da nossa pesquisa foi o projeto, a implementação e a avaliação da ferramenta nAPOLI que é disponibilizada na web livremente para que outros pesquisadores possam utilizar. A ferramenta e suas funcionalidades foram previamente descritas no capítulo 2 e não entraremos em detalhes novamente nesse capítulo. A organização desse capítulo é a seguinte: iniciamos apresentando um comparativo de prós e contras entre nossa ferramenta e as demais apresentadas na Seção 1.4, em seguida descrevemos um estudo de caso utilizado para uma avaliação preliminar da ferramenta e demostração de sua utilidade em diversos casos de uso e finalizamos estabelecendo um comparativo dos resultados levantados com a ajuda do nAPOLI com resultados experimentais.

O estudo de caso tem como objetivo principal descrever como a ferramenta pode ser utilizada para responder às questões propostas na Metodologia (casos de uso) e quais conclusões um usuário pode ter ao utilizar o nAPOLI. Assim, é possível ilustrar como essa ferramenta se mostra proficiente no estudo e análise de padrões de interações proteína-ligante.

Os resultados desse estudo de caso estão disponíveis como exemplo no próprio site do nAPOLI e podem ser acessados através da seção *Results* (*Resultados*) da ferramenta. Na página de ajuda dessa seção (*How to*) há um link[1] para esse estudo de caso. Adicionalmente, esses resultados foram apresentados no X-Meeting em 2014 conforme artigo em anexo a essa dissertação, mas, infelizmente, os organizadores do evento não disponibilizaram os anais do evento. Dessa forma o artigo não está indexado.

[1] http://www.napoli.dcc.ufmg.br/results/consult/cdk2_example/

3.1 Comparativo entre as ferramentas para análise de interações proteína-ligante

As tabelas a seguir resumem as análises de prós e contras de cada ferramenta apresentada neste trabalho. A Tabela 3.1, apresenta um resumo das funcionalidades que cada ferramenta apresenta, enquanto que a Tabela 3.2 apresenta os tipos de interações que cada ferramenta possui.

3.2 Base de dados

O conjunto de dados que escolhemos consiste de uma mesma proteína em complexo com diversos ligantes. A proteína escolhida é a CDK2 (Ciclin-dependent kinase 2) humana e os ligantes são diversos inibidores desenvolvidos a partir de um único composto (identificador PDB X02). Esse ligante foi descoberto através de high-throughput screening (HTS) que é uma técnica em que testes biológicos e químicos são realizados em larga escala [Schonbrunn et al., 2013]. A partir desse experimento os autores descobriram que esse ligante é um inibidor para a CDK2 humana.

Em seguida, os autores modificaram sistematicamente o ligante descoberto e produziram 95 análogos que foram sintetizados e avaliados quanto às suas capacidades inibitórias. O objetivo desse estudo foi propor novos inibidores que tivessem uma capacidade inibitória maior que o ligante X02. Assim, destes 95 compostos, 35 complexos CDK2-ligante tiveram suas estruturas resolvidas através de cristalografia de raio-X. Essas estruturas foram, então, depositadas no Protein Data Bank (PDB). A partir desse trabalho, pesquisamos as estruturas PDB depositadas por eles e encontramos mais 38 estruturas relacionadas que também haviam sido depositadas pelo mesmo autor posteriormente. Entretanto, nesse último caso, o trabalho até este momento ainda não foi publicado e não tivemos acesso aos resultados.

Dessa forma, no total, obtivemos 73 estruturas consistindo de uma mesma proteína (CDK2 humana) em complexo com diferentes ligantes.

3.3 Processamento

Nosso primeiro passo após adquirirmos as 73 estruturas, foi processá-las a fim de produzir todos os dados necessários para realizar nossas análises. Para tanto, realizamos todas as etapas descritas na Metodologia que consiste em computar os tipos

Tabela 3.1. Resumo das funcionalidades de cada ferramenta.

Funcionalidades	Ferramentas					
	LIGPLOT	LigPlot+	Sting	CREDO	GIANT	nAPOLI
Upload de arquivos ou uso de arquivos locais	✓	✓	✓	✓	✓	
Informação sobre acessibilidade ao solvente	✓	✓	✓			
Pontes de hidrogênio intermediadas por H_2O	✓	✓	✓	✓		
Interações entre resíduos que não interagem diretamente com o sítio de ligação	✓	✓	✓			
Interações entre resíduos para a proteína toda			✓			
Parâmetros personalizáveis	✓	✓	✓			✓
Funciona em qualquer navegador	∗	∗		✓		✓
Não requer JAVA instalado	✓			✓		✓
Relaciona informações das proteínas com outras bases de dados				✓		
Não requer atualizações na base de dados para disponibilizar novas estruturas	✓	✓	✓		✓	✓
Busca por ligantes similares				✓		
Requisição via *web service*				✓		
Descreve os modos de ligação (preferências espaciais)					✓	
Comparações de interações em larga escala (inúmeros complexos)		†				✓
Descreve os padrões de interação em larga escala (inúmeros complexos)						✓
Técnicas de visualização de dados para analisar padrões de interação						✓

∗ Não é uma aplicação *web*, então, não requer nenhum navegador.
† Existe uma limitação relacionada ao número de estruturas que se pode analisar simultaneamente no LigPlot+.

atômicos, calcular as interações (utilizamos as distâncias padrão do nAPOLI) e os grafos representando as interações Proteína-ligante, agrupar os ligantes, entre outros.

Realizando o agrupamento obtivemos 7 grupos de ligante similares (Figura 3.1). Então, a análise foi realizada utilizando o conjunto completo de ligantes e cada agru-

Tabela 3.2. Tipos de interações disponíveis em cada ferramenta.

Interações	Ferramentas					
	LIGPLOT	LigPlot+	Sting	CREDO	GIANT*	nAPOLI
Ponte de hidrogênio	✓	✓	✓	✓		✓
Interação hidrofóbica	✓	✓	✓	✓		✓
Empilhamento aromático			✓	✓		✓
Interação repulsiva			✓	✓		✓
Ponte salina			✓	✓		✓
Interação π-cátion				✓		
Complexo metálico				✓		
Interações de van der Waals				✓		
Colisões de van der Waals				✓		
Interações covalentes				✓		
Ponte de halogênio				✓		

$*$ GIANT define apenas que os átomos estão em contato, o tipo da interação em si não é definido.

pamento. Isso porque um padrão global pode não emergir já que os ligantes são bem diversos.

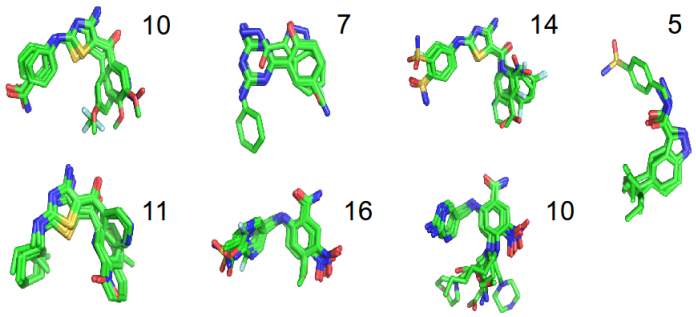


Figura 3.1. Grupos de ligantes obtidos pelo nAPOLI através de uma abordagem automática. O número sobre cada imagem mostra o número de ligantes em cada grupo. Os ligantes são sobrepostos para ajudarem a revelar as similaridades entre as moléculas de cada grupo.

3.4 Análises de padrões proteína-ligante

Nessa seção, descreveremos um exemplo de como diversas análises podem ser executadas através do exemplo do nosso estudo de caso.

3.4.1 Quais são as possíveis interações que cada ligante pode estabelecer e como eles interagem com a proteína?

Para responder estas perguntas utilizamos as seções *Dataset summary* (*Resumo do conjunto de dados*) e *Atoms x Interactions* (*Correlação entre número de átomos e número de interações*) localizada em *Graphical analysis* (*Análises gráficas*).

Através da tabela da seção *Dataset summary*, percebemos que o conjunto de ligantes é bem diverso, contendo desde ligantes com apenas 13 átomos até ligantes maiores contendo 30 átomos. O mesmo pode ser dito para as interações, que variam desde ligantes que estabelecem apenas 2 interações até ligantes que estabelecem 14 interações. Isto pode ser feito facilmente clicando sobre o cabeçalho destas respectivas colunas e ordenando-as de forma crescente ou decrescente (Figura 3.2(c)).

Ressalta-se que todos os três ligantes (3QL8, 3RNI, 3RK7) que estabelecem apenas 2 interações, fizeram-nas por meio de pontes de hidrogênio (Figura 3.2(b)), o que corrobora com as constatações de Schonbrunn et al. [2013]. Para encontrar estes resultados, basta ordenar a coluna *# Protein-ligand interactions* (*Número de interações proteína-ligante*) em ordem decrescente e então selecionar a opção *View details on protein-ligand interactions* (*Ver detalhes sobre interações proteína-ligante*) (Figura 3.2(a)).

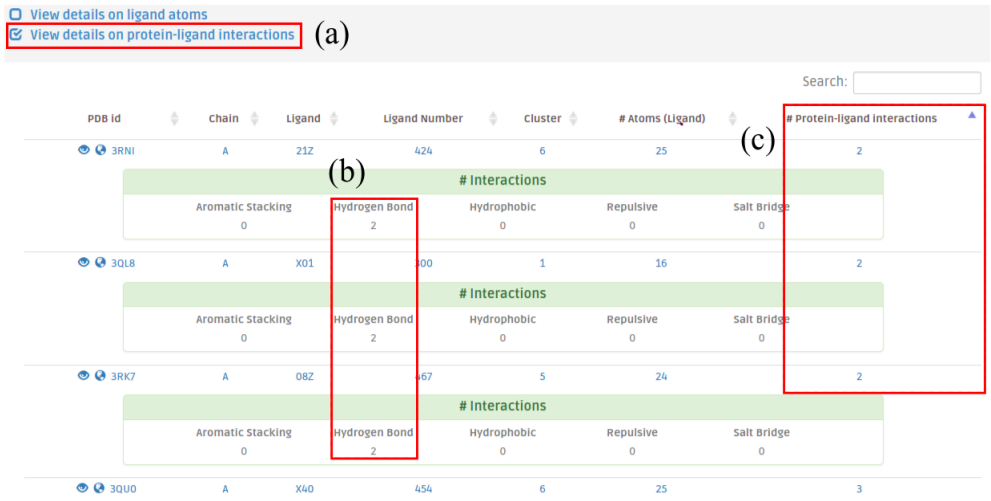


Figura 3.2. Exemplo sobre como utilizar a tabela *Dataset summary*. Para descobrir qual o menor número de interações estabelecidas basta clicar em *View details on protein-ligand interactions* (a) para mostrar os painéis de interações e ordenar a coluna *# Protein-ligand interactions* em ordem crescente (c). Neste exemplo, o menor número de interações é 2 e todas são pontes de hidrogênio.

Além disso, dado o número de átomos disponíveis e o número de interações de

fato estabelecidas podemos concluir que existe uma grande quantidade de átomos que poderiam teoricamente interagir com a proteína e não o fazem.

A partir da seção *Atoms x Interactions*, constatamos que esse conjunto de dados apresentou uma fraca correlação entre o número de átomos aceptores e doadores e o número de pontes de hidrogênio. O mesmo ocorreu entre o número de átomos hidrofóbicos e o número de interações hidrofóbicas. Constatamos também que não existem empilhamentos aromáticos nessa base de dados, mesmo que todos os ligantes tenham esse tipo de átomo. Esses gráficos também mostram os contatos não usados tendo em vista que há muito mais átomos do que interações em cada ponto (Figura 2.15).

3.4.2 Qual é a frequência de cada tipo de átomo no conjunto de ligantes interagindo?

Através da sessão *Atoms type (all data)* (*Tipos de átomos (todos os dados)*) localizada em *Graphical analysis* (*Análises gráficas*), constatamos que todos os ligantes apresentam pelo menos 5 átomos aromáticos. Destes, 39,7% apresentaram 12 átomos, 28,8% apresentaram 17 átomos, 15,1% apresentaram 11 átomos e apenas 11% apresentaram 10 ou menos átomos aromáticos. Logo, cerca de 89% dos ligantes apresentam 11 ou mais átomos aromáticos.

Em todos os ligantes foram encontrados átomos aceptores (Figura 2.13) e átomos doadores. Sendo que 80,8% dos ligantes apresentam entre 4 e 7 átomos aceptores. 95,9% dos ligantes apresentam entre 2 e 4 átomos doadores.

Os dados mostram também que o sítio de ligação para a CDK2 não é muito compatível com átomos carregados. Isto porque apenas 8,2% dos ligantes apresentam 1 átomo positivamente carregado e 8,2% deles apresentam 1 ou 2 átomos negativamente carregado – 8,2% equivale a 6 ligantes.

Em relação à parte apolar, constatou-se que 71,2% dos ligantes apresentam 4 ou menos átomos hidrofóbicos. De 73 ligantes apenas um deles não possui nenhum átomo desse tipo. Em contraste, dois ligantes apresentaram 11 átomos hidrofóbicos.

Entretanto, alguns ligantes também apresentaram átomos que não foram classificados em nenhum tipo pelo Pmapper. Chamamos estes átomos de átomos não classificados (*Unrated atoms*). 54,8% dos ligantes apresentaram 1 átomo não classificado e apenas 5,5% apresentaram 2 átomos desse tipo. Por outro lado, isso significa que 39,7% dos ligantes tiveram todos os seus átomos classificados em algum tipo.

Em resumo, os átomos que foram sempre representados no conjunto de ligantes estudados são os aromáticos, doadores e aceptores. Note que apesar desse relatório ser

simples, é útil para inferir padrões de reconhecimento molecular quando um conjunto de ligantes volumoso está interagindo com uma proteína.

3.4.3 Qual é a frequência de cada tipo de interação no conjunto de ligantes interagindo?

A partir da análise da tabela disponível em *Dataset summary* (*Resumo do conjunto de dados*) sobre as interações estabelecidas entre a CDK2 e seus ligantes, verificamos que 10,9% dos ligantes estabeleceram 4 interações com o receptor, 24,6% apresentaram 5 interações e 16,3% apresentaram 6 interações. Além disso, como mencionado na Seção 3.4.1, todas as moléculas estabeleceram pelo menos duas interações, que nesse caso foram pontes de hidrogênio.

Utilizando os gráficos da seção *Interactions type (all data)* (*Tipos de interações*) (Figura 3.3), vimos facilmente que nenhum ligante estabeleceu empilhamento aromático com a CDK2. Nesse conjunto de ligantes, há sempre pelo menos 1 ponte de hidrogênio, mas 84,9% dos ligantes apresentaram pelo menos 3 interações desse tipo. Como esperado 93,2% dos ligantes não estabeleceram nenhuma interação repulsiva com a CDK2. Porém, quando ela ocorreu foram estabelecidas 1 (2,7%), 2 (2,7%) ou 3 (1,4%) interações. De forma similar, 91,8% dos ligantes não apresentaram nenhuma ponte salina. Sendo que quando ela ocorreu foram estabelecidas 1 (2,7%), 2 (4,1%) e 3 (1,4%) interações. Finalmente, quanto às interações hidrofóbicas, encontramos desde ligantes que estabelecem 11 interações até ligantes que não estabeleceram nenhuma interação desse tipo. 71,2% dos ligantes apresentaram entre 0 e 3 interações hidrofóbicas: 0 (21,9%), 1 (11%), 2 (26%) e 3 (12,3%).

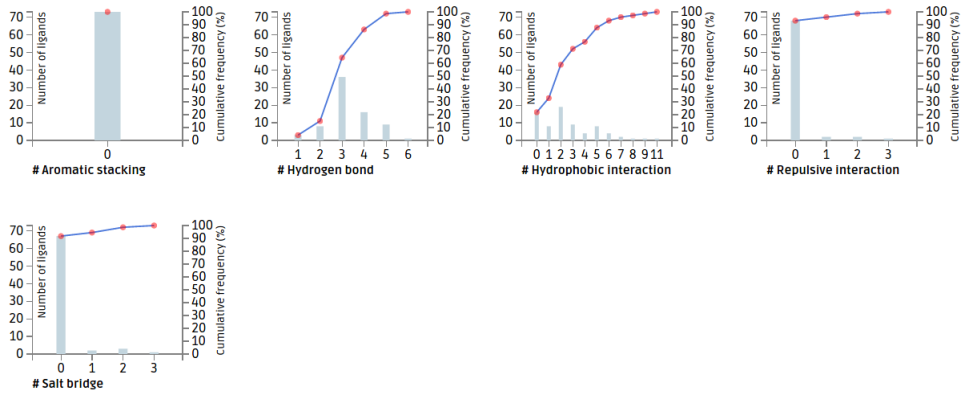


Figura 3.3. Gráficos de barra mostrando a frequência do número de interações para cada tipo.

Apesar de não ter sido encontrado nenhum empilhamento aromático e a grande maioria dos ligantes não terem apresentado pontes salinas e interações repulsivas para esse conjunto de dados, ressalta-se que em outras bases de dados tais contatos podem ser essenciais para a manutenção da ligação.

3.4.4 Quais são os resíduos interagindo com os ligantes?

Com base na tabela disponível na seção *Interactions by residues* (*Interações por resíduos*), observamos 20 diferentes resíduos interagindo com o conjunto completo de ligantes. Apenas um deles (Leucina 83 - LEU83) interagiu com todos os ligantes (100% de frequência) (Figura 2.16). Nesse resíduo, o oxigênio da carboxila sempre interagiu através de pontes de hidrogênio com os ligantes. Enquanto que o nitrogênio da amina estabeleceu pontes de hidrogênio com 57,5% dos ligantes. O segundo resíduo que mais frequentemente interagiu com os ligantes foi o Glutamato 81 (GLU81), que interagiu com 83,56% das moléculas. Em todas as vezes, o oxigênio do grupo carboxila desse resíduo estabeleceu pontes de hidrogênio com os ligantes. Já o átomo CD1 da Leucina 134 (LEU134), estabeleceu interações hidrofóbicas com 57,5% dos ligantes. Depois desse átomo, todos os outros resíduos não apresentaram interações muito conservadas (menos de 50% dos ligantes).

3.4.5 Existem grupos de ligantes similares?

Em um conjunto contendo 73 ligantes é esperado a existência de grupos de moléculas similares entre si e diferentes das demais moléculas. E de fato isso ocorre nesse conjunto de dados. Vale ressaltar que o nAPOLI aplica um algoritmo que realiza o agrupamento dos ligantes automaticamente (ver Seção 2.9) com o objetivo de simplificar as análises para os usuários. Além disso, todas análises disponíveis para conjunto global de dados também se encontra disponível para cada grupo. Para permitir a análise comparativa de diferentes grupos de ligantes, desenvolvemos também um tipo de gráfico voltado especialmente para esse fim, que são os gráficos de barra agrupada. Estes gráficos estão disponíveis tanto em *Atoms type (clusters)* (*Tipos de átomos (grupos)*) e *Interactions type (clusters)* (*Tipos de interações (grupos)*) da seção *Graphical analysis* (*Análises gráficas*).

Como mencionado na Seção 3.3, nAPOLI encontrou 7 grupos de ligantes (Figura 3.1). Constatamos que de fato estes grupos são bastante dissimilares entre si. Através dos gráficos de barras agrupadas, verificamos que os grupos 4 e 7 apresentam um número de átomos aceptores muito diferentes. Enquanto que o grupo 4 apresenta entre

2 a 4 átomos desse tipo, o grupo 7 apresenta 6 ou mais. Por outro lado, o número de átomos doadores parece ser mais bem distribuído entre os vários grupos do que os átomos aceptores. Isto porque os grupos de ligantes apresentam uma alta frequência de 3 átomos doadores. Átomos positivos e negativos também possuem uma distribuição homogênea para os diferentes grupos e são em sua maioria iguais a 0. Da mesma forma que os átomos aceptores, a distribuição de número de átomos hidrofóbicos também é bem diversa. É possível ver que o grupo 4 apresenta muito mais átomos hidrofóbicos do que os demais agrupamentos, enquanto que o grupo 6 é um dos grupos com menor número de átomos desse tipo. Por fim, os grupos 5 e 6 são os grupos com mais átomos aromáticos, enquanto que o grupo 4 possui poucos átomos desse tipo.

A respeito das interações, constatou-se que o grupo 4 é o grupo com mais interações hidrofóbicas, que era algo esperado uma vez que esse grupo também é o grupo com mais átomos hidrofóbicos. Empilhamentos aromáticos, interações repulsivas e pontes salinas possuem uma grande similaridade entre os grupos uma vez que quase todos eles não possuem estas interações. Já em relação às ponte de hidrogênio, verificou-se que essas interações são mais bem distribuídas do que as demais. A maioria dos grupos possuem entre 3 e 4 pontes de hidrogênio (Figura 2.14).

É difícil verificar a qualidade destes agrupamentos. Apesar de Schonbrunn et al. [2013] apresentarem grupos diferentes de moléculas, deve-se ressaltar que os grupos gerados por eles não foram obtidos a partir do agrupamento de moléculas similares e sim através de modificações feitas nas moléculas a fim de obter uma atividade inibitória maior e, de fato, os grupos deles não são comparáveis aos nossos.

3.5 Comparações dos achados do nAPOLI com resultados experimentais

Como citado na Seção 3.2, Schonbrunn et al. [2013] descobriu o composto 2-(allylamino)-4-aminothiazol-5-yl-(phenyl)methanone (identificador PDB X02) através de high-throughput screening (HTS). Esse composto atua como inibidor ligando-se ao sítio de ligação de ATP da CDK2 com um IC_{50} igual a 15μM. Um IC_{50} representa qual a concentração de uma dada molécula é necessária para inibir 50% do alvo (receptor). Essa medida reflete a eficiência de uma substância química em inibir um receptor, de forma que quanto menor o valor de um IC_{50} mais eficiente é o inibidor. Vale ressaltar que essa medida depende da concentração do substrato que é utilizado no experimento. Portanto, se os experimentos não são realizados sob as mesmas condições, é impossível utilizar o IC_{50} como medida comparativa [Yung-Chi e Prusoff, 1973].

Os autores resolveram a estrutura do complexo CDK2 com o ligante X02 e depositaram essa estrutura no PDB (identificador 3QQK). Com essa estrutura, os autores confirmaram que o composto de fato se liga no sítio de ATP da CDK2 através de pontes de hidrogênio com a região de dobradiça (*hinge region*), que é composta pelos resíduos GLU81-LEU83. A partir desse composto e com o objetivo de produzir um inibidor com uma eficiência maior, os autores sintetizaram 95 análogos, mantendo, porém, o grupo *aminothiazol* responsável por interagir com os resíduos da região de dobradiça. A síntese envolveu a substituição sistemática apenas dos grupos funcionais *phenyl* e *allyl*.

A partir da tabela disponível em *Interactions by residues* (*Interações por resíduos*), verificamos que os resíduos mais conservados em relação às interações entre a CDK2 e seus inibidores são os resíduos LEU83 (100%) e GLU81 (83,56%) ambos envolvidos em pontes de hidrogênio. A LEU83 estabelece pontes de hidrogênio com 73 ligantes a partir do oxigênio da carboxila. Enquanto que o nitrogênio da amina estabeleceu pontes de hidrogênio com 57,5% dos ligantes, isto é, com 42 ligantes. Já o GLU81 estabeleceu pontes de hidrogênio com 61 ligantes através do oxigênio da carboxila.

Modificações realizadas no grupo *phenyl*, resultaram em 14 análogos (6 estruturas resolvidas: 3R8V, 3QTQ, 3R8U, 3S00, 2S0O e 3R8Z). Estas estruturas foram agrupadas através de nossa metodologia no grupo (*cluster*) 3. Analisando a conservação de resíduos interagindo com os ligantes para esse grupo, constatou-se que os resíduos mais conservados são **GLU81** (100%), **LEU83** (100%), ILE10 (72,73%), **LYS33** (63,64%), LEU134 (63,64%) e **PHE82** (54,55%). Os resíduos em negrito estão de acordo com o que os autores mencionam no artigo para esse grupo. Vale ressaltar que, no artigo, os autores não mencionam nada sobre as interações hidrofóbicas desse grupo, pois eles focaram apenas nas pontes de hidrogênio que essas estruturas estabeleceram. Porém, na análise do ligante X02 (estrutura 3QQK) os autores definem explicitamente as interações hidrofóbicas estabelecidas com os resíduos ILE10 e LEU134, indicando, assim, que esses dois resíduos provavelmente interagem tal como nossa ferramenta apontou. Por outro lado, não encontramos nenhum indicativo de interações hidrofóbicas estabelecidas pelo resíduo PHE82 no artigo. Além disso, no artigo os autores também mencionam que o resíduo ASP145 estabelece uma interação de ponte de hidrogênio intermediada por água com o ligante. Portanto, como desconsideramos moléculas de água nesta versão de nossa ferramenta, não fomos capazes de encontrar esta interação.

Os autores também mencionam as interações entre os análogos 3QU0 e 3QXP e a CDK2, que foram os análogos que apresentaram a maior atividade inibitória graças à substituição do grupo alil (*allyl*) por um grupo para-fenil sulfonamida. Essa substituição aumentou significativamente o estabelecimento de uma rede de pontes de

Atom	Type of interaction	# Ligands with which it interacts
	POS81 (GLU81) : GLU81	Total: 11 (100.00%)
O	Hydrogen bond	11
	POS83 (LEU83) : LEU83	Total: 11 (100.00%)
N	Hydrogen bond	3
O	Hydrogen bond	11
	POS10 (ILE10) : ILE10	Total: 8 (72.73%)
CD1	Hydrophobic	4
CG2	Hydrophobic	5
	POS33 (LYS33) : LYS33	Total: 7 (63.64%)
CE	Hydrophobic	5
NZ	Hydrogen bond	7
	POS134 (LEU134) : LEU134	Total: 7 (63.64%)
CD1	Hydrophobic	6
CD2	Hydrophobic	1
	POS82 (PHE82) : PHE82	Total: 6 (54.55%)
CE2	Hydrophobic	6
CZ	Hydrophobic	5

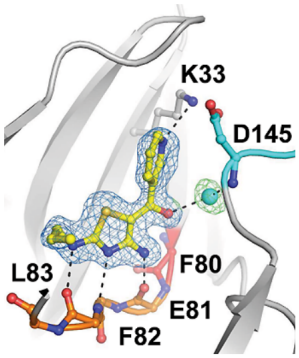


Figura 3.4. Tabela disponível em Interações por resíduo mostrando os resíduos mais frequentes do grupo 3. Retângulos azuis destacam os resíduos que estão em conformidade com o que Schonbrunn et al. [2013] mencionam sobre as estruturas 3R8V, 3QTQ, 3R8U, 3S00, 2S0O e 3R8Z. A figura à direita, mostra o sítio de ligação da CDK2 em complexo com o ligante X35 (PDB 3QTQ) e os resíduos que interagem com esse ligante (Fonte: Schonbrunn et al. [2013]).

hidrogênio com os resíduos **ASP86**, **ASP145**, **LYS33**, **GLU81**, **LEU83** e GLN85. Os 5 resíduos em negrito também foram encontrados como os resíduos mais conservados no grupo 6, que foi onde estas estruturas foram agrupadas através de nossa abordagem: GLU81 (100%), LEU83 (100%), ASP86 (78,57%), ASP145 (35,71%) e LYS33 (28,57%). No artigo os autores mencionam que os resíduos HIS84, GLN85, LYS89 e ASP145 estabelecem interações de ponte de hidrogênio intermediada por água com o ligante. Isso justifica, portanto, o motivo de não termos encontrado essas interações, uma vez que desconsideramos moléculas de água nesta versão de nossa ferramenta.

Quando analisamos os análogos 3QTZ, 3QTU, 3QTX e 3RPV, também gerados pela substituição do grupo alil por um grupo sulfonamida, obtivemos resultados similares. Todos eles foram automaticamente inseridos no grupo 6, exceto a estrutura 3QTX. Os autores apontam os resíduos **ASN132**, **ASP145**, GLY13, **LYS33**, **PHE80**, **VAL18** e **GLN131** como resíduos interagindo com os compostos análogos. A partir do grupo 6, encontramos 6 de 7 resíduos (resíduos em negrito) com uma alta frequência de acordo com a tabela disponível em *Interactions by residues* (*Interações por resíduos*) do nAPOLI. Sendo que, no artigo, os autores mencionam que o resíduo GLY13 estabelece interações hidrofóbicas entre o carbono-α (CA) e os ligantes. Assim,

como em nossa metodologia não classificamos o CA como sendo do tipo hidrofóbico, não foi possível identificar as interações desse resíduo.

Já no grupo 5, no qual a estrutura 3QTX foi agrupada, encontramos apenas os resíduos LYS33 e VAL18 como resíduos interagindo. Os demais resíduos, segundo os autores, interagem através de interações hidrofóbicas e interações-π com dois oxigênios do ligante X43. Essas interações, sobretudo a interação hidrofóbica com um átomo polar, não foram modelados em nossa ferramenta e, portanto, não fomos capazes de encontrá-las. Além disso, ressalta-se que o fato de termos encontrado dois grupos distintos para essas estruturas, deve-se à abordagem que utilizamos para agrupar os ligantes de forma automática. Abordagem essa muito diferente daquela aplicada por Schonbrunn et al. [2013], uma vez que os autores produziram esses grupos a partir de modificações sistemáticas nos grupos funcionais dos compostos.

Dessa forma, concluímos que o nAPOLI foi capaz de apontar resíduos importantes que compõem os padrões de interações entre um receptor e um conjunto de ligantes segundo a literatura. É importante levar em consideração que Schonbrunn et al. [2013] estavam interessados no potencial inibitório dos ligantes e esse não foi o foco de nosso trabalho. Entretanto, a ferramenta mostrou ser efetiva na detecção de resíduos que são importantes para o reconhecimento molecular e, portanto, é promissora e proficiente para isso. Encontramos diversos resultados que estão de acordo com as interações que Schonbrunn et al. [2013] notaram experimentalmente e que nAPOLI foi capaz de selecioná-las automaticamente. De fato, houve também algumas diferenças devido à forma como modelamos o problema. Podemos citar, por exemplo, as interações de pontes de hidrogênio intermediadas por água que não fomos capazes de encontrar. Entretanto, como o nAPOLI foi desenvolvido em módulos, essas diferenças podem ser facilmente solucionadas a partir de modificações na modelagem de nossa ferramenta.

Também concluímos que os padrões de interação, como esperado, não são simples de serem inferidos e podem ser conservados em grupos de moléculas ao invés de serem conservados como um único padrão de reconhecimento molecular no conjunto completo de dados. A abordagem de agrupamento implementada no nAPOLI mostrou ser muito útil para auxiliar na descoberta do que é conservado ou não em um subconjunto de ligantes e quais interações eles estabelecem em comum com o receptor.

Apesar de existirem outras ferramentas para a análise de interações proteína-ligante, nenhuma delas apresenta um grande conjunto de dados estatísticos que possam ser analisados de forma visual e de modo interativo ou mesmo que facilitem e permitam a realização de análises comparativas das interações como é feito no nAPOLI.

Capítulo 4

Conclusão

Propomos neste trabalho um conjunto de interfaces visuais e interativas para revelar padrões de interações proteína-ligante, contribuindo, assim, para a elucidação dos processos envolvidos no reconhecimento molecular. Desenvolvemos, nesse trabalho, nAPOLI (Analysis of PrOtein Ligand INteractions), uma ferramenta *web* que permite que sejam realizadas análises estatísticas, visuais e interativas em larga escala para o estudo dos padrões de interações proteína-ligante. Essa ferramenta permite o uso de quaisquer conjuntos de dados, sejam aqueles constituídos por uma proteína e vários ligantes, várias proteínas e um ligante ou várias proteínas e vários ligantes diferentes.

Realizamos um estudo de caso que consistia de 73 inibidores para a proteína humana CDK2 (Ciclin-dependent kinase 2) e o nAPOLI mostrou ser muito útil, pois ajudou na descoberta do que era conservado ou não nas interações proteína-ligante. Através da ferramenta, fomos capazes de confirmar resultados experimentais da literatura. Também encontramos alguns resultados diferentes, porém, ressalta-se que essas diferenças podem ser facilmente solucionadas através de modificações na modelagem da ferramenta. Além disso, nossa ferramenta também mostra-se capaz de encontrar e propor resultados não mencionados antes pela literatura e que podem ser confirmados por estudos posteriores. Portanto, esperamos que a ferramenta possa ser bastante útil para a comunidade científica.

Apesar da existência de outras ferramentas interessantes para analisar interações proteína-ligante, nenhuma delas apresenta análises estatísticas, visuais e interativas em larga escala o que dificulta a comparação do nosso trabalho com as demais. Em geral, essas ferramentas não oferecem mecanismos para comparações de complexos em larga escala através de dados estatísticos, tais como a frequência de resíduos que mais interagem com os ligantes, tipos de átomos ou interações. Dessa forma, o usuário deve realizar suas análises de forma manual.

4.1 Trabalhos futuros

Neste trabalho, nosso objetivo foi projetar, implementar e avaliar uma nova ferramenta *web* que permita a análise de padrões de interações proteína-ligante. Através dessa ferramenta, foi possível, inclusive, descrever e confirmar resultados da literatura. Porém, a fim de melhorar nossa ferramenta, diversas outras funcionalidades podem ser incluídas. A seguir serão definidos alguns trabalhos que podem ser desenvolvidos para aperfeiçoar nossa ferramenta e nossa metodologia:

- Permitir que o usuário utilize uma base de dados própria através do *upload* de arquivos. Isso é útil, sobretudo, para pesquisas envolvendo ancoragem molecular em que são gerados diversos complexos proteína-ligante considerando várias conformações do ligante ou da proteína;
- Incluir íons, aminoácidos não padrão e moléculas de água no estudo das interações proteína-ligante;
- Permitir que o usuário faça pesquisas de PDBs apenas utilizando uma sequência de aminoácidos;
- Permitir que o usuário escolha qual abordagem para cálculo de contatos ele deseja utilizar;
- Permitir o uso de diferentes estados de protonação do aminoácido e do ligante, e variados valores de pH do ambiente;
- Incluir novos tipos de interações, tais como pontes de halogênio;
- Possibilitar que o usuário realize o *download* dos resultados;
- Permitir a pesquisa de estruturas que contenham um determinado ligante ou ligantes similares;
- Oferecer novas funcionalidades e visualizações para a ferramenta *web*

Uma importante perspectiva da ferramenta é que ela deixe de ser apenas descritiva e passe a ser preditiva. O que queremos dizer é que o nAPOLI foi inicialmente concebido para ajudar na análise em larga escala de padrões de reconhecimento molecular em complexos resolvidos experimentalmente e nos quais os ligantes são obviamente conhecidos. De posse desse conhecimento, pretendemos que seja possível incorporar técnicas de triagem virtual e de ancoragem molecular na predição de novos ligantes que possam ser ancorados às proteínas e ter suas interações preditas analisadas pela

ferramenta. Felizmente, esse trabalho será executado uma vez que acabamos de ter esse projeto aceito para ser executado no Doutorado do Programa de Pós-Graduação em Bioinformática. Por fim, temos a importante perspectiva de empregar a ferramenta em outros estudos de caso dentre os quais mencionamos prioritariamente o da previsão de ligantes neutralizadores de Ricina. Trata-se de um projeto em rede já financiado pela CAPES e através do qual poderemos ter validações experimentais para as predições realizadas através de colaborações com experimentalistas envolvidos na equipe do projeto.

Referências Bibliográficas

Altschul, S.; Gish, W.; Miller, W.; Myers, E. e Lipman, D. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Altschul, S.; Madden, T.; Schäffer, A.; Zhang, J.; Zhang, Z.; Miller, W. e Lipman, D. (1997). Gapped BLAST and textscPSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402.

Andrabi, M.; Nagao, C.; Mizuguchi, K. e Ahmad, S. (2009). *Bioinformatics Approaches for Analysis of Protein–Ligand Interactions*, pp. 267–299. John Wiley & Sons, Inc.

Attias, H. (1999). Inferring parameters and structure of latent variable models by variational bayes. Em *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, pp. 21–30, São Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Ballester, P. e Richards, W. (2007). Ultrafast shape recognition to search compound databases for similar molecular shapes. *Journal of Computational Chemistry*, 28(10):1711–1723.

Bashton, M.; Nobeli, I. e Thornton, J. (2008). PROCOGNATE: A cognate ligand domain mapping for enzymes. *Nucleic Acids Research*, 36(SUPPL. 1):D618–D622.

Bateman, A.; Martin, M.; O'Donovan, C.; Magrane, M.; Apweiler, R.; Alpi, E.; Antunes, R.; Arganiska, J.; Bely, B.; Bingley, M.; Bonilla, C.; Britto, R.; Bursteinas, B.; Chavali, G.; Cibrian-Uhalte, E.; Da Silva, A.; De Giorgi, M.; Dogan, T.; Fazzini, F.; Gane, P.; Castro, L.; Garmiri, P.; Hatton-Ellis, E.; Hieta, R.; Huntley, R.; Legge, D.; Liu, W.; Luo, J.; MacDougall, A.; Mutowo, P.; Nightingale, A.; Orchard, S.; Pichler, K.; Poggioli, D.; Pundir, S.; Pureza, L.; Qi, G.; Rosanoff, S.; Saidi, R.; Sawford, T.; Shypitsyna, A.; Turner, E.; Volynkin, V.; Wardell, T.; Watkins, X.; Zellner, H.; Cowley, A.; Figueira, L.; Li, W.; McWilliam, H.; Lopez, R.; Xenarios, I.; Bougueleret, L.; Bridge, A.; Poux, S.; Redaschi, N.; Aimo, L.; Argoud-Puy, G.; Auchincloss,

A.; Axelsen, K.; Bansal, P.; Baratin, D.; Blatter, M.-C.; Boeckmann, B.; Bolleman, J.; Boutet, E.; Breuza, L.; Casal-Casas, C.; De Castro, E.; Coudert, E.; Cuche, B.; Doche, M.; Dornevil, D.; Duvaud, S.; Estreicher, A.; Famiglietti, L.; Feuermann, M.; Gasteiger, E.; Gehant, S.; Gerritsen, V.; Gos, A.; Gruaz-Gumowski, N.; Hinz, U.; Hulo, C.; Jungo, F.; Keller, G.; Lara, V.; Lemercier, P.; Lieberherr, D.; Lombardot, T.; Martin, X.; Masson, P.; Morgat, A.; Neto, T.; Nouspikel, N.; Paesano, S.; Pedruzzi, I.; Pilbout, S.; Pozzato, M.; Pruess, M.; Rivoire, C.; Roechert, B.; Schneider, M.; Sigrist, C.; Sonesson, K.; Staehli, S.; Stutz, A.; Sundaram, S.; Tognolli, M.; Verbregue, L.; Veuthey, A.-L.; Wu, C.; Arighi, C.; Arminski, L.; Chen, C.; Chen, Y.; Garavelli, J.; Huang, H.; Laiho, K.; McGarvey, P.; Natale, D.; Suzek, B.; Vinayaka, C.; Wang, Q.; Wang, Y.; Yeh, L.-S.; Yerramalla, M. e Zhang, J. (2015). UniProt: A hub for protein information. *Nucleic Acids Research*, 43(D1):D204–D212.

Berger, C.; Weber-Bornhauser, S.; Eggenberger, J.; Hanes, J.; Plückthun, A. e Bosshard, H. (1999). Antigen recognition by conformational selection. *FEBS Letters*, 450(1-2):149–153.

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

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA. ISBN 0387310738.

Boehr, D. e Wright, P. (2008). How do proteins interact? *Science*, 320(5882):1429–1430.

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

Branden, C. e Tooze, J. (1999). *Introduction to Protein Structure*. Nova York: Garland Science, 2 edição. 410p.

Brewerton, S. (2008). The use of protein-ligand interaction fingerprints in docking. *Current Opinion in Drug Discovery and Development*, 11(3):356–364.

Bricogne, G.; Blanc, E.; Brandl, M.; Flensburg, C.; Keller, P.; Paciorek, W.; Roversi, P.; Sharff, A.; Smart, O.; Vonrhein, C. e T.O, W. (2011). BUSTER. Versão 2.10.1. Cambridge, United Kingdom: Global Phasing Ltd. https://www.globalphasing.com/buster/wiki/index.cgi?. Acessado em: 13/06/2015.

Böhm, H.-J. (2005). Prediction of non-bonded interactions in drug design. Em *Protein-Ligand Interactions*, pp. 3–20. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA.

Caceres, R.; Pauli, I.; Timmers, L. e de Azevedo Jr., W. (2008). Molecular recognition models: A challenge to overcome. *Current Drug Targets*, 9(12):1077–1083.

CGAL (1995). Cgal, Computational Geometry Algorithms Library. http://www.cgal.org.

Chartrand, G.; Lesniak, L. e Zhang, P. (2010). *Graphs & Digraphs, Fifth Edition*. A Chapman & Hall book. Taylor & Francis.

Chemaxon (2014). Pharmacophore perception - pmapper. Disponível em: `https://docs.chemaxon.com/display/CD/Pharmacophore+perception+-+PMapper`. Acessado em: 31/05/2015.

Clark, M.; Cramer, R. D. e Van Opdenbosch, N. (1989). Validation of the general purpose tripos 5.2 force field. *Journal of Computational Chemistry*, 10(8):982–1012.

Corey, Robert B.; Pauling, L. (1953). Molecular models of amino acids, peptides, and proteins. *Review of Scientific Instruments*, 24.

Csermely, P.; Palotai, R. e Nussinov, R. (2010). Induced fit, conformational selection and independent dynamic segments: An extended view of binding events. *Trends in Biochemical Sciences*, 35(10):539–546.

da 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.

Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A. e Laufer, J. (1992). Description of several chemical structure file formats used by computer programs developed at molecular design limited. *Journal of Chemical Information and Computer Sciences*, 32(3):244–255.

Davey, C.; Sargent, D.; Luger, K.; Maeder, A. e Richmond, T. (2002). Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *Journal of Molecular Biology*, 319(5):1097–1113.

Daylight (2011). Daylight theory manual. Disponível em: `http://www.daylight.com/dayhtml/doc/theory/`. Acessado em: 29/05/2015.

de Beer, T. A. P.; Berka, K.; Thornton, J. M. e Laskowski, R. A. (2014). PDBsum additions. *Nucleic Acids Research*, 42(D1):D292–D296.

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

de Melo, R. C.; Gomide, J. S.; Dias, P. S.; Meira Jr, W. e Santoro, M. M. (2007b). Mining structural signatures of proteins. Em *XXII Simpósio Brasileiro de Banco de Dados. III Workshop em Algoritmos e Aplicações de Mineração de Dados. João Pessoa-PB.*

de Melo, R. C.; Lopes, C.; Fernandes Jr, F. A.; da Silveira, C. H.; Santoro, M. M.; Carceroni, R. L.; Meira Jr, W. e Araújo Ade, A. (2006). A contact map matching approach to protein structure similarity analysis. *Genet. Mol. Res*, 5(2):284–308.

Deng, Z.; Chuaqui, C. e Singh, J. (2004). Structural interaction fingerprint (sift): A novel method for analyzing three-dimensional protein-ligand binding interactions. *Journal of Medicinal Chemistry*, 47(2):337–344.

Dokholyan, N. V. (2012). *Computational Modeling of Biological Systems: From Molecules to Pathways.* Biological and Medical Physics, Biomedical Engineering. Springer.

Downs, G. M. e Barnard, J. M. (2003). *Clustering Methods and Their Uses in Computational Chemistry*, pp. 1–40. John Wiley & Sons, Inc.

Dunn, M. F. (2010). *Protein–Ligand Interactions: General Description.* John Wiley & Sons, Ltd.

Ernst, J.; Clubb, R.; Zhou, H.-X.; Gronenborn, A. e Clore, G. (1995). Demonstration of positionally disordered water within a protein hydrophobic cavity by nmr. *Science*, 267(5205):1813–1817.

Few, S. (2009). *Now You See It: Simple Visualization Techniques for Quantitative Analysis.* Oakland (California): Analytics Press, 1 edição. 327 p.

Finn, R.; Bateman, A.; Clements, J.; Coggill, P.; Eberhardt, R.; Eddy, S.; Heger, A.; Hetherington, K.; Holm, L.; Mistry, J.; Sonnhammer, E.; Tate, J. e Punta, M. (2014). Pfam: The protein families database. *Nucleic Acids Research*, 42(D1):D222–D230.

Foote, J. e Milstein, C. (1994). Conformational isomerism and the diversity of antibodies. *Proceedings of the National Academy of Sciences of the United States of America*, 91(22):10370–10374.

Forli, S. (2015). Charting a path to success in virtual screening. *Molecules*, 20(10):18732–18758.

Gehlenborg, N.; O'Donoghue, S.; Baliga, N.; Goesmann, A.; Hibbs, M.; Kitano, H.; Kohlbacher, O.; Neuweger, H.; Schneider, R.; Tenenbaum, D. e Gavin, A.-C. (2010). Visualization of omics data for systems biology. *Nature Methods*, 7(3 SUPPL.):S56–S68.

Gellman, S. H. (1997). Introduction: molecular recognition. *Chemical reviews*, 97(5):1231–1232.

Gonçalves-Almeida, V. M. (2011). *HydroPaCe: uma metodologia para análise de inibição cruzada em serino proteases através de centroides de regiões hidrofóbicas*. Tese, Universidade Federal de Minas Gerais, Universidade Federal de Minas Gerais, Belo Horizonte.

Grunenberg, J. (2011). Complexity in molecular recognition. *Physical Chemistry Chemical Physics*, 13(21):10136–10146.

Hendlich, M.; Bergner, A.; Günther, J. e Klebe, G. (2003). Relibase: Design and development of a database for comprehensive analysis of protein–ligand interactions. *Journal of Molecular Biology*, 326(2):607 – 620.

Humphrey, W.; Dalke, A. e Schulten, K. (1996). VMD - Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38.

James, L.; Roversi, P. e Tawfik, D. (2003). Antibody multispecificity mediated by conformational diversity. *Science*, 299(5611):1362–1367.

Jayaram, B. e Jain, T. (2004). The role of water in protein-dna recognition. *Annual Review of Biophysics and Biomolecular Structure*, 33:343–361.

Kasahara, K. e Kinoshita, K. (2014). GIANT: pattern analysis of molecular interactions in 3d structures of protein-small ligand complexes. *BMC Bioinformatics*, 15:12.

Kasahara, K.; Shirota, M. e Kinoshita, K. (2013). Comprehensive classification and diversity assessment of atomic contacts in protein-small ligand interactions. *Journal of Chemical Information and Modeling*, 53(1):241–248.

Kelley, L.; Gardner, S. e Sutcliffe, M. (1996). An automated approach for clustering an ensemble of nmr-derived protein structures into conformationally related subfamilies. *Protein Engineering*, 9(11):1063–1065.

Kessel, A. e Ben-Tal, N. (2010). *Introduction to Proteins: Structure, Function, and Motion*. Chapman & Hall/CRC Mathematical and Computational Biology. Londres: CRC Press. ISBN 9781439810729. 654 p.

Kleywegt, G. J. (2007). Crystallographic refinement of ligand complexes. *Acta Crystallographica Section D*, 63(1):94–100.

Koshland, D. E. J. (1958). Application of a theory of enzyme specificity to protein synthesis. *Proc Natl Acad Sci U S A*, 44.

Lahti, J.; Tang, G.; Capriotti, E.; Liu, T. e Altman, R. (2012). Bioinformatics and variability in drug response: A protein structural perspective. *Journal of the Royal Society Interface*, 9(72):1409–1437.

Laskowski, R. e Swindells, M. (2011). Ligplot+: Multiple ligand-protein interaction diagrams for drug discovery. *Journal of Chemical Information and Modeling*, 51(10):2778–2786.

Lavecchia, A. e Giovanni, C. (2013). Virtual screening strategies in drug discovery: A critical review. *Current Medicinal Chemistry*, 20(23):2839–2860.

Leach, A. R. e Gillet, V. J. (2007). *An Introduction to Chemoinformatics*. Springer Publishing Company, Incorporated.

Levy, Y. e Onuchic, J. (2006). Water mediation in protein folding and molecular recognition. *Annual Review of Biophysics and Biomolecular Structure*, 35:389–415.

Mackinnon, J.; Gallastegui, N.; Osguthorpe, D.; Hagler, A. e Estébanez-Perpiñá, E. (2014). Allosteric mechanisms of nuclear receptors: Insights from computational simulations. *Molecular and Cellular Endocrinology*, 393(1-2):75–82.

Mancini, A. L.; Higa, R. H.; Oliveira, A.; Dominiquini, F.; Kuser, P. R.; Yamagishi, M. E. B.; Togawa, R. C. e 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.

Marcou, G. e Rognan, D. (2007). Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *Journal of Chemical Information and Modeling*, 47(1):195–207.

Martínez, L.; Andreani, R. e Martínez, J. (2007). Convergent algorithms for protein structural alignment. *BMC Bioinformatics*, 8.

McDonald, I. K. e Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *Journal of Molecular Biology*, 238(5):777–793.

Metzler, D. E. (2003). *Biochemistry: The Chemical Reactions of Living Cells*, volume 1 e 2. Nova York: Academic Press, 2 edição. 1973 p.

Meyer, E. (1992). Internal water molecules and h-bonding in biological macromolecules: A review of structural features with functional implications. *Protein Science*, 1(12):1543–1562.

Murzin, A.; Brenner, S.; Hubbard, T. e 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. e Cox, M. M. (2014). *Princípios de Bioquímica de Lehninger*. Porto Alegre: Artmed, 6 edição. 1328 p.

O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T. e Hutchison, G. R. (2011). Open babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(10).

Open Babel (2012). The open babel package, versão 2.3.1. http://openbabel.org.

Park, S. e Saven, J. (2005). Statistical and molecular dynamics studies of buried waters in globular proteins. *Proteins: Structure, Function and Genetics*, 60(3):450–463.

Pearson, W. e Lipman, D. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 85(8):2444–2448.

Pickett, S. (2005). The biophore concept. Em *Protein-Ligand Interactions*, pp. 73–105. Wiley-VCH Verlag GmbH & Co. KGaA.

Pires, D. E.; de Melo-Minardi, R. C.; dos Santos, M. A.; da Silveira, C. H.; Santoro, M. M. e Meira, W. (2011). Cutoff scanning matrix (csm): structural classification and function prediction by protein inter-residue distance patterns. *BMC genomics*, 12(Suppl 4):S12.

Pires, D. E. V.; De Melo-Minardi, R. C.; Da Silveira, C. H.; Campos, F. F. e Meira Jr., W. (2013). Acsm: Noise-free graph-based signatures to large-scale receptor-based ligand prediction. *Bioinformatics*, 29(7):855–861.

Poupon, A. (2004). Voronoi and voronoi-related tessellations in studies of protein structure and interaction. *Current Opinion in Structural Biology*, 14(2):233–241.

Rego, N. e Koes, D. (2015). 3Dmol.js: molecular visualization with WebGL. *Bioinformatics*, 31(8):1322–1324.

Robson, B. e Vaithiligam, A. (2009). *Drug Gold and Data Dragons: Myths and Realities of Data Mining in the Pharmaceutical Industry*, pp. 25–85. John Wiley & Sons, Inc.

Rupp, B. (2009). *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*. Taylor & Francis Group. ISBN 9781134064199.

Saraiya, P.; North, C. e Duca, K. (2005). An insight-based methodology for evaluating bioinformatics visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 11(4):443–456.

Schonbrunn, E.; Betzi, S.; Alam, R.; Martin, M. P.; Becker, A.; Han, H.; Francis, R.; Chakrasali, R.; Jakkaraj, S.; Kazi, A. et al. (2013). Development of highly potent and selective diaminothiazole inhibitors of cyclin-dependent kinases. *Journal of medicinal chemistry*, 56(10):3768–3782.

Schreyer, A. e Blundell, T. (2009). Credo: A protein–ligand interaction database for drug discovery. *Chemical Biology & Drug Design*, 73(2):157–167.

Schrödinger, LLC (2010). The PyMOL molecular graphics system, version 1.7r6.

Schroeder, M.; Gilbert, D.; Van Helden, J. e Noy, P. (2001). Approaches to visualisation in bioinformatics: From dendrograms to space explorer. *Information Sciences*, 139(1-2):19–57.

Schwabe, J. (1997). The role of water in protein-dna interactions. *Current Opinion in Structural Biology*, 7(1):126–134.

Shatsky, M.; Nussinov, R. e Wolfson, H. J. (2004). A method for simultaneous alignment of multiple protein structures. *Proteins: Structure, Function, and Bioinformatics*, 56(1):143–156.

Silveira, S.; Fassio, A.; Goncalves-Almeida, V.; de Lima, E.; Barcelos, Y.; Aburjaile, F.; Rodrigues, L.; Meira Jr, W. e de Melo-Minardi, R. (2014). Vermont: Visualizing mutations and their effects on protein physicochemical and topological property conservation. *BMC Proceedings*, 8(Suppl 2):S4. ISSN 1753-6561.

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

Sreenivasan, U. e Axelsen, P. (1992). Buried water in homologous serine proteases. *Biochemistry*, 31(51):12785–12791.

Stryer, L.; Tymoczko, J. L. e Berg, J. M. (2004). *Bioquímica*. Rio de Janeiro: Guanabara Koogan, 5 edição. 1097 p.

Strömbergsson, H. e Kleywegt, G. J. (2009). A chemogenomics view on protein-ligand spaces. *BMC bioinformatics*, 10 Suppl 6:S13.

Tanaka, S. e Scheraga, H. A. (1975). Model of protein folding: inclusion of short-, medium-, and long-range interactions. *Proceedings of the National Academy of Sciences*, 72(10):3802–3806.

Todeschini, R.; Consonni, V.; Mannhold, R.; Kubinyi, H. e Timmerman, H. (2008). *Handbook of Molecular Descriptors*. Methods and Principles in Medicinal Chemistry. Wiley.

Tsai, C.-J.; Kumar, S.; Ma, B. e Nussinov, R. (1999). Folding funnels, binding funnels, and protein function. *Protein Science*, 8(6):1181–1190.

Tsai, C.-J.; Ma, B.; Sham, Y. Y.; Kumar, S. e Nussinov, R. (2001). Structured disorder and conformational selection. *Proteins: Structure, Function, and Bioinformatics*, 44(4):418–427.

Verli, H. (2014). *Bioinformática: da Biologia à Flexibilidade Molecular (2014)*. Porto Alegre, Brasil, 1 edição. Disponível em: http://www.ufrgs.br/bioinfo/ebook/.

Wallace, A.; Laskowski, R. e Thornton, J. (1995). Ligplot: A program to generate schematic diagrams of protein-ligand interactions. *Protein Engineering*, 8(2):127–134.

Wang, L.; Xie, Z.; Wipf, P. e Xie, X.-Q. (2011). Residue preference mapping of ligand fragments in the protein data bank. *Journal of Chemical Information and Modeling*, 51(4):807–815.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

Wong, B. (2012). Points of view: visualizing biological data. *Nature Methods*, 9(12):1131.

Yung-Chi, C. e Prusoff, W. (1973). Relationship between the inhibition constant (ki) and the concentration of inhibitor which causes 50 per cent inhibition (i50) of an enzymatic reaction. *Biochemical Pharmacology*, 22(23):3099–3108.

Zaki, M. e Meira, W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Zvelebil, M. e Baum, J. O. (2008). *Understanding bioinformatics*. Nova York: Garland Science, 1 edição. 772 p.

Apêndice A

Publicação

Artigo intitulado “Visual and interactive strategies to reveal patterns of protein-ligand interaction” submetido no congresso ISCB-Latin America x-Meeting on Bioinformatics with BSB and SoiBio, 2014. Apresentamos a ferramenta *web* nAPOLI e os resultados de nosso estudo de caso, que foi um conjunto de dados constituído da proteína CDK2 (Ciclin-dependent kinase 2) humana em complexo com 73 ligantes diferentes.

Através deste trabalho fomos convidados também para uma apresentação oral no evento. Infelizmente, os organizadores desse congresso não disponibilizaram os anais do evento e, portanto, o artigo não está indexado.

RESEARCH

Visual and interactive strategies to reveal patterns of protein-ligand interactions

Alexandre V Fassio[1,2], Sabrina A Silveira[1] and Raquel C de Melo-Minardi[1*]

*Correspondence:
raquelcm@dcc.ufmg.br
[1]Department of Computer Science, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6.627, Instituto de Ciências Exatas, 31270-901, Belo Horizonte, Brazil
Full list of author information is available at the end of the article

Abstract

Background: Molecular recognition plays an important role in biological systems and is observed between receptor-ligand, antigen-antibody, DNA-protein, RNA-ribossome, etc. Molecular recognition is a phenomenon of organization very difficult to predict or design even for small molecules. Due to its remarkable importance, molecular recognition was studied under different perspectives in Bioinformatics. Several studies focused on seeking patterns of molecular recognition on datasets consisting of a specific receptor and multiple ligands or vice versa by using varied data mining techniques. The analytical process is extremely toilsome as an expert has to assay each of the patterns carefully and they can be very voluminous. Therefore, this paper proposes a quite different approach which aims at being more easy and intuitive. The use of images to represent information is becoming more and more appreciated for the benefits it can bring to science by providing a powerful means both to make sense of data and to communicate. Data visualization has in recent years become an established area of study in academia and is increasingly being used in biological data visualization. In this paper, we propose visual and interactive strategies to depict the types of interactions established between a protein and its ligands and materialized its strategies into a tool called nAPOLI (Analysis of PrOtein Ligand Interactions). Particularly, to show an example of use of the tool, we are going to focus on a case study of an important family of enzymes: the Cyclin Dependent Kinases II (CDK2).

Results: Our results indicate that nAPOLI is very useful as it shed some light on the problem of discovering what is conserved or not in a set of ligands and interactions they establish with a receptor. We were able to confirm some experimental results from literature and other points were raised by the tool but not mentioned before and still have to be confirmed by further studies. Despite the existence of other interesting tools to analyze protein-ligand contacts none presents a large scale statistical, visual and interactive analysis of the contacts data as nAPOLI.

Conclusion: We proposed a set of visual and interactive interfaces to reveal patterns of protein-ligand interactions and to shed some light on how the molecular recognition process occurs. We implemented a prototype of the system and used a particular dataset that consisted of a set of 73 ligands inhibiting the human CDK2 obtaining interesting results in agreement with literature and others to be further investigated.

Keywords: interactions; contacts; molecular recognition; ligand; receptor; CDK2; protein; information visualization; visual interfaces; interactive

Background

A ligand is a substance, commonly a small molecule, that forms a complex with a biomolecule serving a biological purpose. We usually say that the ligand is a molecule that binds a binding site on a target or receptor protein. This binding occurs by intermolecular non-covalent bonding such as van der Waals forces, ionic or hydrogen bonds, hydrophobic forces and π-π interactions and the functional state of the receptor is determined by its conformational state. The association is usually reversible and the strength of binding is called affinity, which depends upon not only direct interactions but also solvent effects that can play a dominant indirect role in driving the non-covalent interactions in solution.

The term molecular recognition refers to the specific interactions that occurs between two or more molecules. As a matter of fact, ligand and receptor involved in molecular recognition must exhibit molecular complementarity [1]. Molecular recognition plays an important role in biological systems and is observed between receptor-ligand, antigen-antibody, DNA-protein, RNA-ribossome, etc. It can be static or dynamic. In the first case, there is a host and a guest molecule like a key and a keyhole and the recognition sites must be very specific for the guest molecules. In the case of the dynamic molecular recognition the binding of a first guest to a first binding site can affect the association constant of a second guest molecule. According to [2], molecular recognition is a phenomenon of organization very difficult to predict or design even for small molecules.

Due to its remarkable importance molecular recognition was studied under different perspectives in Bioinformatics. Several studies focused on seeking patterns of molecular recognition on datasets consisting of a specific receptor and multiple ligands or vice versa. Some of these approaches are graph-based as graphs are a straightforward approach to model the molecular recognition phenomenon. Nodes are the atoms of proteins and small molecules and edges are the chemical interactions that can be established between them. One of the most promising approaches are the algorithms to find frequent subgraphs [3, 4, 5, 6] in the protein-ligand complexes.

Nevertheless, these approaches have an important drawback concerning the exponential number of patterns that can be generated in the mining process. This makes extremely toilsome the analytical process as an expert has to assay each of the patterns carefully. Therefore, this paper proposes a quite different approach which aims at being more easy and intuitive. The use of images to represent information is becoming more and more appreciated for the benefits it can bring to science by providing a powerful means both to make sense of data and to communicate. Data visualization has in recent years become an established area of study in academia and is increasingly being used in biological data visualization [7]. In this paper, we propose visual strategies to depict the types of interactions established between protein and its ligands. Furthermore, we implemented these visual strategies in a prototype tool called nAPOLI (Analysis of PrOtein Ligand Interactions) which is available for user evaluation[1].

[1] http://www.dcc.ufmg.br/~alexandrefassio/napoli

Methods

In this section, we will explain how nAPOLI is organized, its functionalities and some methodological aspects on how the data was computed. It was developed to answer some basic but important questions about how a different set of ligands interact to a common receptor:

1 **What are the possible *interactions each ligand* can establish and how does it in fact interact with the protein?** Is there a correlation between the number of atoms and the number of interactions established? We wanted to know what types of atoms each ligand present in terms of what types of interactions each atom can do. Therefore, we classified atoms as hydrophobic, positive, negative, donor, acceptor, aromatic and / or sulfur. Hence we would like to show a quantitative report on how many atoms of each type exist and how many in fact interact with the receptor.

2 **What is the frequency of each *type of atom in the whole set of interacting ligands*?** Here our objective was to present a profile of the set of ligands in terms of the distribution of the occurrences of atom types.

3 **What is the frequency of each type of *interaction in the whole set of interacting ligands*?** We wanted to establish a molecular recognition profile in terms of the distribution of the interactions between the complete set of ligands and the receptor.

4 **What are the *receptor interacting residues*?** And what is the frequency each residue interacts with the varied set of ligands? When a residue interacts, what type of interactions it establishes and with what frequency? What are the interacting atoms?

5 **Are there *groups of similar ligands*?** And how the above questions can be answered by each cluster? Are the clusters similar or dissimilar?

The rest of this section is organized as follows: first, we introduce the dataset we used to test the proposed visual and interactive strategies. Then, we present the methodology to compute the protein-ligand contacts, the main type of data our system presents. After that, we explain how we modeled and computed protein-ligand graphs and subgraphs. Finally, we introduce and discuss the visual and interactive strategies we proposed to make possible the analysis of the patterns of protein-ligand interactions.

Protein-ligands dataset

We chose a model dataset composed of a varied set of ligands bound to a identical protein. We departed from the work of [8] on the human CDK2 where authors describe the development of potent diaminothiazole inhibitors from a single compound discovered by high throughput screening. By searching in the Protein Data Bank (PDB) web site, we found 73 related PDB entries which are the ones used in this study. They are all composed by the same human CDK2 and a very diverse set of ligands.

Computing protein-ligand contacts

We used a cutoff-independent approach to geometrically compute probable amino acid interactions at atomic level, which we also mapped to residue level. For each

protein, we generated a Voronoi diagram followed by its Delaunay tessellation [9, 10] and, using both distance and physicochemical properties described in [11], we classified contacts into one class.

Protein-ligand contacts were computed as follows:

1 Protein atoms were classified according to its chemical characteristics as being hydrophobic, positive, negative, donor, acceptor, aromatic and / or sulfur. This classification was done manually according to our previous work [12, 13, 14, 15, 16].
2 Ligand atoms were classified in the same classes according to the software PMapper [17, 18] into the same classes as the protein atoms.
3 Delaunay tessellation was computed to determine hypothetical contacts using CGAL software library [19].
4 A distance criteria was used to detect if a pair of atoms of certain types is in contact: $4.5\mathring{A}$ for two hydrophobic atoms (hydrophobic interaction), $6\mathring{A}$ for a positive and a negative atom (salt bridge), $6\mathring{A}$ for to atoms with the same charge (repulsive contact), $3.2\mathring{A}$ for a donor and an acceptor atom (hydrogen bond) and $8\mathring{A}$ for two aromatic atoms (aromatic stacking).

Clustering the ligands

As we imagined the set of ligands would be very diverse and a global pattern could not emerge by having a global view of the data, we decided to cluster the compounds and perform separate analysis for each cluster instead of only a general view.

First of all, we converted each ligand from the PDB format to mol using Open Babel. The second step was to generate fingerprints for each ligand using the GenerateMD from Chemaxon with the following default parameters: *descriptor type*: "CF" (Chemical Fingerprint), *length*: 1,024, *bit-count*: 2, *pattern length*: 7 and *input/output format*: decimal-format.

The ligands were then clustered using the software Ward from Chemaxon. The first step of the process consisted of running the software with the parameter -K to apply Kelley method with aims at automatically selecting the number of clusters. Thus, in this first run the following parameters were used: *fingerprint size*: 1,024, *generate id* (-g) and *Kelley method* (-K). We got 7 clusters as the best number of groups. Then we run the method again with the following parameters: *fingerprint size*: 1,024, *generate id* (-g) and *number of clusters*: 7. The clusters of ligands are shown in Figure 11.

Computing protein-ligand graphs and subgraphs

We generated a set of graphs to represent the protein-ligand data. For each protein-ligand interface we designed a graph where the nodes are protein and ligand atoms that are in contact (according to its chemical characteristics and distance criteria) and the edges are the interactions established among these atoms.

These graphs were represented in Graphml [20] which is a XML based comprehensive file format that allows the representation of application-specific attribute data. Here, the type of each atom and interaction, atom index in PDB and node color were the so called application-specific attributes. The graphs can be accessed in the *Protein-ligand interactions by ligands* section of nAPOLI (blue nodes are

protein atoms and the orange ones are ligand atoms) and users can see at a glance the amount of protein and ligand atoms involved in the interactions, the types of interactions and the graph topology in a complimentary manner to the structure of the protein-ligand complex. An example of graph (for protein-ligand interface of PDB id 3R8L) is provided in Figure 8.

The modeling of interactions as graphs allows us to search frequent subgraphs in the protein-ligand interface, which can reveal non obvious relevant patterns. As a preliminary step toward the subgraph mining, for each graph, we computed its set of induced subgraphs using the algorithm [21], which does not take into account the properties of nodes and edges. Therefore, we proposed a summary for subgraphs that considers their number and type of interactions or edges. For each subgraph, we computed a vector of length 7 in which position 1 has the PDB id, position 2 has the number of interactions and positions 3 to 7 store how these interactions are distributed among its possible types. The detailed subgraphs dataset is available as a table in the *Protein-ligand detailed subgraphs* section of our tool.

Also, we summarized these data in the *Protein-ligand subgraphs summary* section of nAPOLI. In this table, each line is a type of subgraph and, in a similar manner to *Protein-ligand detailed subgraphs*, the first column is the number of interactions, columns 2 to 6 show the types of such interactions and the last column shows how many times this type of subgraph appears in the whole dataset.

In both tables we used a heatmap representation in which color intensity is a pre-attentive attribute that encodes the number of a certain type of interaction. The aim of this representation was to bring forth an overview of the complete subgraph data, evidencing trends and exceptions in subgraph interactions across the CDK2 dataset (73 PDB ids). The mining of frequent subgraphs is a future work.

The plots of graphs as well as its subgraphs were computed using graph-tool [22] software.

Contacts statistics and visual strategies

To answer the first stated question in the section Methods about the possible interactions each ligand can establish, we designed an interactive table with summary of the number of ligand atoms and the number of protein-ligand interactions for each ligand (PDB id). By clicking on the total number of atoms or on the total number of interactions, users can obtain a sub table with the number of each type of atoms or interactions. We also have a checkbox where the users can open all the atom details panel or all the interactions details panel. With these options, users can compare the number of contacts established by the different ligands with the common receptor. If the user wants to see the contacts in the protein structure, it is possible to click on an eye icon in each line of the table and get a GLMol plug in with the protein in cartoon and the ligand and the protein interacting residues in sticks colored in CPK as shown in Figure 2. Thus users can interact with the structure by rotating and zooming the binding site. The set of PDB files were all superposed in a way that they are all in the same orientation making easier users comparisons. There is also a link to the PDB web site where users can find the original data about each PDB file.

Furthermore, concerning the correlation between the number of atoms of each type and the number of established interactions, we developed a scatterplot with this data were users can visually verify the correlation as shown in Figure 4.

Concerning the second question about the frequency of atom types in the ligand dataset we designed a visual interactive interface where users can find the distribution number of atoms of a specific type by the number of ligands in a histogram shown in Figure 3. In the bar chart, users can analyze how a set of values is distributed in an interval as well as contrast how two or more set of values are distributed. This graph depicts many interesting visual patterns as the spread of the data, the center of the set of values, the shape of the distribution: if it is curved or flat, if it is upward or downward, if it has a single peak or multiple peaks, if it is symmetrical or skewed to left or right, if it has some concentrations or gaps or even outliers. Each of these patterns can show an interesting feature of the data. Furthermore, we added a line chart and a second y-axis representing the cumulative values of percentage frequencies, what is commonly known as Pareto chart. It is a very interesting visual representation to highlight the most important among a set of values. Notice that the chart is interactive and users can obtain details (precise values) on demand by passing the mouse over the bars or the points of the line. This interface is also very useful to answer the first posed question as it graphically displays similar information as the dataset summary.

To complement this analysis, we provided pie charts to depict the part-to-whole data relationship as shown in Figure 3. From our point of view, it is less precise than bar chart in this analysis, but can give a complementary view of the set of slices that are prominent. The pie chart is also very interactive and users can not only obtain details on demand but also select a set of slices that sum up a certain percentage of the data going from bottom to up (departing from lower values) and top to down (departing from higher values). It can be very useful when users want to know what slices are responsible for, for instance, 50% of the data.

The third question about the frequency of contact types was solved in a quite identical solution as second question with Pareto and pie charts.

Regarding the fourth question, about the most important interacting residues, we presented two solutions. Firstly, we designed an interactive color coded table as depicted in Figures 5 and 6. In this table, we have basically the columns: Atom, Type of interaction and Number of interacting ligands. However the lines with these basic columns can be grouped by header lines with the following columns: Residue and Percentage of interacting ligands. This grouping lines are colored according to a code that goes from blue (lower values) to orange (higher values). It seems a very simple visual representation by it has a remarkable feature: the various sorting possibilities and the possibility of search in any of its fields. By clicking on each column header, users can order the data by the basic columns: Atom, Type of interaction and Number of interacting ligands. By clicking on any of the grouping lines (the colored ones) users can order the data by residue name or sequential number as well as by the percentage of interacting ligands. For instance, if users sort the data by this last feature they can readily find the most important residues that most frequently interact with the set of ligands as shown in Figure 5. By searching for a specific set of characters, users can filter out the data very easily.

For instance, if they type ALA, only Alanine residues will remain in the table. If they type, CB, only β carbons will remain.

The second solution to this question is a matrix where lines are the ligands (PDB files) and the columns are possible residue atoms as shown in Figure 7. In the cells we present the residue in contact, the type of interaction and the ligand atom. It is a very simple way to show the data and to filter out unnecessary data. For instance, if users type "repuls", they are going to get only ligands that present at least one repulsive contact with the protein. If they type HIS, they are going to get ligands that interact with an Histidine. In addition, users can see these data visually in graph plots where blue nodes present protein residues name and number and atom and orange nodes depict ligand atoms as shown in Figure 8.

Regarding the fourth question about the set of statistics explained above for clusters of similar molecules, we implemented the same set of analysis previously mentioned but segmented by clusters.

Results and discussion

In this section we present some conclusions users can reach using nAPOLI in answering the questions we posed in the section Methods.

What are the possible interactions each ligand can establish and how does it in fact interact with the protein?

The biggest ligand of the CDK2 dataset has 30 atoms and the smallest, 13. It is easy to see this by ordering the lines of dataset summary in table presented in Figure 1. There is a ligand that does 19 interactions with the protein and three ligands established only 2 contacts with the receptor, both of them hydrogen bonds. Thus we can say there are a lot of "unused" atoms that could be in contact but that are not.

Furthermore, Figure 4 shows that, for this particular dataset, we have a weak correlation between the number of atoms of acceptors and donors and the number of hydrogen bonds and between the number of hydrophobic atoms and the number of hydrophobic interactions. This graph also shows the "unused" atoms as there are much more atoms than interactions for each point.

What is the frequency of each type of atom in the whole set of interacting ligands?

The analysis of the ligands showed that all the molecules present at least 5 aromatic atoms. 38.67% of the ligands presented 12 aromatic atoms, 30.67% presented 17 and 14.67% presented 11. Only 10.67% presents 10 or less aromatic atoms what means about 90% of the ligands present more than 11 aromatic atoms.

There is no ligand without donor or acceptor atoms. 71.33% of the ligands present between 4 and 7 acceptor atoms. 96% of the ligands present between 2 and 4 donors.

Data show that the CDK2 binding site is not very compatible with charged ligands. Only 10% of the ligands present 1 positively charged atom and 8% of them present 1 or 2 negatively charged atoms.

Concerning the apolar portion of ligands, 72% of the ligands present 4 or less hydrophobic atoms but there are two ligands with 11 hydrophobic atoms and on the other side one with 0 atoms of this type.

Some atoms are not classified into one of these types according to PMapper and they are what we call unrated. 40% of the ligands do not present an unrated atom, 54.67% present only 1 and only four ligands present 2 unrated atoms.

In summary, the types of atoms that are always represented in the set of studied ligands are the aromatic, donors and acceptors. Notice that this report is very simple but useful to infer molecular recognition patterns when a voluminous set of ligands is interacting with a receptor.

What is the frequency of each type of interaction in the whole set of interacting ligands?

The analysis of the contacts established between enzyme and its ligands shows that 11% of the molecules presents 4 interactions with the receptor, 26% presents 5 contacts with the receptor and 16%, 6 contacts. All the molecules establish at least 2 contacts and in these cases they are hydrogen bonds.

There are no aromatic stackings. There is always at least one hydrogen bond, but 85% of the molecules present at least 3 contacts of this type. 93.33% of the ligands present no repulsive contacts at all what is expected. When it happened, we observed 1 (1.33% of the ligands), 2 (2.67%) and 3 (2.67%). 92% of the ligands present no salt bridges. When it happened we observed 1 (2.67% of the ligands), 2 (4%) and 3 (1.33%). The number of hydrophobic contacts ranges from 0 to 16. 70.67% of the molecules present between 0 and 3: 0 (22.67%), 1(10.67%), 2 (25.33%) and 3 (12%).

The subgraph data provided in *Protein-ligand subgraphs* (Figures 9 and 10) section of nAPOLI confirms these findings, as one can visually notice that columns associated with hydrophobic and hydrogen bonds have a lot more cells in green than columns associated to salt bridges and repulsive contacts. There are no aromatic interactions, as the last column is in white. Also, in *Protein-ligand subgraphs summary* section of nAPOLI we observe that the most frequent subgraph type (249) is the one involving just one hydrogen bond interaction, followed by the type with one hydrophobic interaction (215) and the ones involving just 2 (147) and 3 (73) hydrophobic interactions respectively. There are 5 types of subgraphs that occur less often (just 1 occurrence for each): the type which involves 1 salt bridge and 1 hydrogen bond; the one with 3 repulsive interactions; 2 salt bridges and 1 hydrogen bond; 1 salt bridge and 3 repulsive interactions; 2 salt bridges and 2 repulsive interactions.

What are the receptor interacting residues?

There are 20 different residues interacting with the whole set of ligands. Only one of them (LEU83) interacts with every ligand (100% of frequency). The carboxyl O always does a hydrogen bond with the ligand. The amine N does a hydrogen bond in 56% of the cases. The second most common interacting residue is GLU81 whose carboxyl O does a hydrogen bond in 84% of the cases. LEU134 CD1 does a hydrophobic interaction with 58.67% of the ligands. After these residues, the others do not present very conserved interactions (less than 50% of the ligands)

Are there groups of similar ligands?

The answer to this question is quite obvious as in a set of 73 ligands we would expect the existence of clusters of homogeneous molecules which would be dissimilar from the rest of clusters. In fact, this happened in this dataset but it is important emphasize that nAPOLI implements an automatic strategy to cluster molecules and to make easy the user analysis. The results of the clustering process can be found in Figure 11. The whole set of analysis is available for the global dataset and for each cluster. There is also an analysis, shown in Figure 12, that makes possible the comparison of the distributions of types of atoms in each cluster of ligands.

Through nAPOLI, we could see that the clusters are quite dissimilar as expected. Clusters 1 and 7 present quite dissimilar number of acceptor atoms: while cluster 1 presents always 6 or more atoms, cluster 7 presents between 2 and 4. It seems that the distribution of donor atoms between the different clusters are much more similar than for the acceptors. Most of the cluster present a high frequency of 3 atoms. The distribution of positive and negative atoms are also very similar for the different clusters and are mostly zero. Hydrophobic atoms are also very dissimilar across the clusters. Notice how cluster 7 present much more hydrophobic atoms than the others. Clusters 4 and 6 present much more aromatic atoms as the others while cluster 7 is a group with none or very few aromatic atoms.

It is difficult to assess the quality of the clusters. Despite [8] presents different clusters of molecules, this clusters were not obtained by grouping similar molecules but they were compiled manually according to modifications done in molecules in order to high its inhibitory activity and, in fact, their clusters are not comparable to ours.

Comparisons to literature experimental results

[8] discovered by high throughput screening the (2-(allylamino)-4-aminothiazol-5-yl-(phenyl)methanone) as an inhibitor of the human CDK2-cyclin A2 complex with an IC_{50} value of 15μM. They also have done a co-crystal structure (3QQK) that confirmed that the compound binds ATP site through hydrogen bonding interactions between the thiazolamine moiety and the hinge region (GLU81-LEU83). The authors designed a set of analogues of the compound such that the hinge-binding functionality of the aminothiazole core remained unchanged while the flanking allyl and phenyl moieties were systematically modified. A total of 95 analogues were synthesized and evaluated using enzymatic assays and crystal structures of 35 CDK2-inhibitor complexes were determined between 1.4 and 2.0 $\AA$ resolution.

As mentioned previously, our dataset is composed by 73 structures because it was enriched with other complexes involving the human CDK. Shonbrunn and colleagues have another work [23] to be published about CDK2 structures but unfortunately, we had not access to it. Using the *Interactions by residues* table, we can verify that the residues that are the most conserved in the interactions between inhibitor and receptor are LEU83 (100%) and GLU81 (84%) both involved in hydrogen bonds. LEU83 does hydrogen bonds with 75 ligands (2 of the PDB files present 2 alternative locations) through the carboxyl O and 42 through amine N. The GLU81 does hydrogen bonds through the carboxyl O with 63 ligands.

Modifications to the phenyl group resulted in 14 analogues (6 solved structures: 3R8V, 3QTQ, 3R8U, 3S00, 2S0O and 3R8Z). These structures were clustered

together (cluster 3) according to our clustering methodology. When we analyze the most conserved interacting residues we get **GLU81** (100%), **LEU83** (100%), ILE10(72.73%), LEU134(63.64%), **LYS33** (63.64%) and **PHE82** (54.55%). GLU81, LEU83, PHE82 and LYS33. Residues in bold are in accordance with what authors mention in the paper for this cluster. However, from our methodology to compute contacts we got that PHE82 does hydrophobic interactions with the ligand and not hydrogen bond as mentioned in their work.

The authors of [8] also mention interactions between sulfonamide analogues (3QU0 and 3QXP) which they found to be the most potent inhibitors due to the addition of a para-phenyl sulfonamide moiety at the allyl site what substantially increased the establishment of a hydrogen bond network with **ASP86**, **ASP145**, **LYS33**, **GLU81**, **LEU83** and GLN85. Five of them can be found as the most frequent interacting when we analyze the table of the cluster 4 where both structures are found: ASP86 is 80% frequent, ASP145 is 33.33%, LYS33 is 26.67%, GLU81 is 100% and LEU83 is 100%.

We got similar results when analyzing the sulfonamide inhibitors with phenyl substituents (3QTZ, 3QTU, 3QTX and 3RPV). Authors mention **ASN132**, **ASP145**, GLY13, **LYS33**, **PHE80**, **VAL18** and **GLN131** as interacting residues. Notice that 6 out of 7 are frequent according to the table of interacting residues of nAPOLI.

In conclusion, nAPOLI was able to point out important residues to compound the pattern of interactions between a receptor and a set of ligands. It is important to stress that in [8] the authors are interested in the inhibitory activity potency what was not the objective when we designed nAPOLI. However, the tool seemed to be effective in pointing out important residues for molecular recognition and we believe it is a promising approach for that. There are many intersections between the interactions authors manually noticed as important and what nAPOLI automatically selected. Some differences do exist but we believe they can be further investigated experimentally.

We also concluded that the interaction pattern is, as expected, not simple to be inferred and can be conserved in groups of molecules instead of in the whole dataset as a unique molecular recognition pattern. As nAPOLI implemented an automatic clustering procedure, it showed to be very useful as it shed some light on the problem of discovering what is conserved or not in a subset of ligands and interactions they establish with the receptor.

Despite the existence of other interesting tools to analyze protein-ligand contacts as LPC [24], the RCSB Ligand Explorer and STING Contacts [25], in special Protein Ligand Contacts, none of these tools present a large scale statistical, visual and interactive analysis of the contacts data as nAPOLI. From our point of view, a model to be able to detect a profile of molecular recognition must know how to deal with a set of diverse (and sometimes voluminous) compounds interacting to a receptor and mine patterns on it.

Conclusions

In this paper, we proposed a set of visual and interactive interfaces to reveal patterns of protein-ligand interactions and to shed some light on how the molecular recognition process occurs. We also implemented a prototype of the system called nAPOLI

(Analysis of PrOtein Ligand INteractions) and used a particular dataset that consists of 73 ligands inhibiting the human Cyclin Dependent Kinases II (CDK2).

nAPOLI showed to be very useful as it helps to discover what is conserved or not in protein-ligand interactions. We were able to confirm experimental results from literature and other points were raised by the tool but not mentioned before and they still have to be confirmed by further studies. Despite the existence of other interesting tools to analyze protein-ligand contacts, to the best of our knowledge none present a large scale statistical, visual and interactive analysis of the contacts data.

Availability and requirements

nAPOLI is available through the link
www.dcc.ufmg.br/~alexandrefassio/napoli

Competing interests
The authors declare that they have no competing interests.

Author's contributions
AF and SAS carried out the laboratory work, analyzed the data, interpreted the results and drafted the manuscript. AF implemented the tool. RCMM contributed to the conception, design and coordination of the study and drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgments
This work was supported by the Brazilian agencies Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), Financiadora de Estudos e Projetos (FINEP) and Pró-Reitoria de Pesquisa da Universidade Federal de Minas Gerais.

Author details
[1]Department of Computer Science, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6.627, Instituto de Ciências Exatas, 31270-901, Belo Horizonte, Brazil. [2]Department of Biochemistry and Immunology, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6.627, Instituto de Ciências Biológicas, 31270-901, Belo Horizonte, Brazil.

References

1. Gellman, S.H.: Introduction: molecular recognition. Chemical reviews **97**(5), 1231–1232 (1997)
2. Grunenberg, J.: Complexity in molecular recognition. Physical Chemistry Chemical Physics **13**(21), 10136–10146 (2011)
3. Yan, X., Han, J.: gspan: Graph-based substructure pattern mining. In: Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference On, pp. 721–724 (2002). IEEE
4. Zou, Z., Li, J., Gao, H., Zhang, S.: Mining frequent subgraph patterns from uncertain graph data. Knowledge and Data Engineering, IEEE Transactions on **22**(9), 1203–1218 (2010)
5. Zhu, F., Yan, X., Han, J., Philip, S.Y.: gprune: a constraint pushing framework for graph pattern mining. In: Advances in Knowledge Discovery and Data Mining, pp. 388–400. Springer, ??? (2007)
6. Wang, W., Wang, C., Zhu, Y., Shi, B., Pei, J., Yan, X., Han, J.: Graphminer: a structural pattern-mining system for large disk-based graph databases and its applications. In: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, pp. 879–881 (2005). ACM
7. Wong, B.: Points of view: visualizing biological data. Nature Methods **9**(1131) (2012)
8. Schonbrunn, E., Betzi, S., Alam, R., Martin, M.P., Becker, A., Han, H., Francis, R., Chakrasali, R., Jakkaraj, S., Kazi, A., *et al.*: Development of highly potent and selective diaminothiazole inhibitors of cyclin-dependent kinases. Journal of medicinal chemistry **56**(10), 3768–3782 (2013)
9. Poupon, A.: Voronoi and voronoi-related tessellations in studies of protein structure and interaction. Current opinion in structural biology **14**(2), 233–241 (2004)
10. Okabe, A., Boots, B., Sugihara, K.: Spatial Tessellations: Concepts and Applications of Voronoi Diagrams. John Wiley & Sons, Inc., New York, NY, USA (1992)
11. Sobolev, V., Sorokine, A., Prilusky, J., Abola, E., Edelman, M.: Automated analysis of interatomic contacts in proteins. Bioinformatics **15**(4), 327–332 (1999)
12. de Melo, R.C., Lopes, C., Fernandes Jr, F.A., da Silveira, C.H., Santoro, M.M., Carceroni, R.L., Meira Jr, W., Araújo Ade, A.: A contact map matching approach to protein structure similarity analysis. Genet. Mol. Res **5**(2), 284–308 (2006)
13. de Melo, R.C., Gomide, J.S., Dias, P.S., Meira Jr, W., Santoro, M.M.: Mining structural signatures of proteins. In: XXII Simpósio Brasileiro de Banco de Dados. III Workshop em Algoritmos e Aplicações de Mineração de Dados. João Pessoa–PB (2007)

14. de Melo, R., Ribeiro, C., Murray, C., Veloso, C., da Silveira, C., Neshich, G., Meira Jr, W., Carceroni, R., Santoro, M.: Finding protein-protein interaction patterns by contact map matching. Genet. Mol. Res **6**(4), 946–963 (2007)
15. da 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.*: 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 (2009)
16. Pires, D.E., de Melo-Minardi, R.C., dos Santos, M.A., da Silveira, C.H., Santoro, M.M., Meira, W.: Cutoff scanning matrix (csm): structural classification and function prediction by protein inter-residue distance patterns. BMC genomics **12**(Suppl 4), 12 (2011)
17. Schneider, G., Lee, M.-L., Stahl, M., Schneider, P.: De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. Journal of Computer-Aided Molecular Design **14**(5), 487–494 (2000)
18. Schneider, G., Clément-Chomienne, O., Hilfiger, L., Schneider, P., Kirsch, S., Böhm, H.-J., Neidhart, W.: Virtual screening for bioactive molecules by evolutionary de novo design. Angewandte Chemie International Edition **39**(22), 4130–4133 (2000)
19. CGAL, Computational Geometry Algorithms Library. http://www.cgal.org
20. Graphml website. Available: http://graphml.graphdrawing.org/. Accessed 2014 Jul 25
21. Wernicke, S.: Efficient detection of network motifs. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) **3**(4), 347–359 (2006)
22. graph-tool website. Available: http://graph-tool.skewed.de/. Accessed 2014 Jul 25
23. Schonbrunn, E., Becker, A., Betzi, S., Alam, R., Han, H., Rawle, F., Katta, V., Jakkaraj, J., Chakrasali, R., Neelam, S., Hook, D., Tash, J., Georg, G.: Structure-guided optimization of novel CDK2 inhibitors discovered by high-throughput screening. To be published
24. Sobolev, V., Sorokine, A., Prilusky, J., Abola, E.E., Edelman, M.: Automated analysis of interatomic contacts in proteins. Bioinformatics **15**(4), 327–332 (1999)
25. Mancini, A.L., Higa, R.H., Oliveira, A., Dominiquini, F., Kuser, P.R., Yamagishi, M.E., Togawa, R.C., Neshich, G.: 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 (2004)

Figures

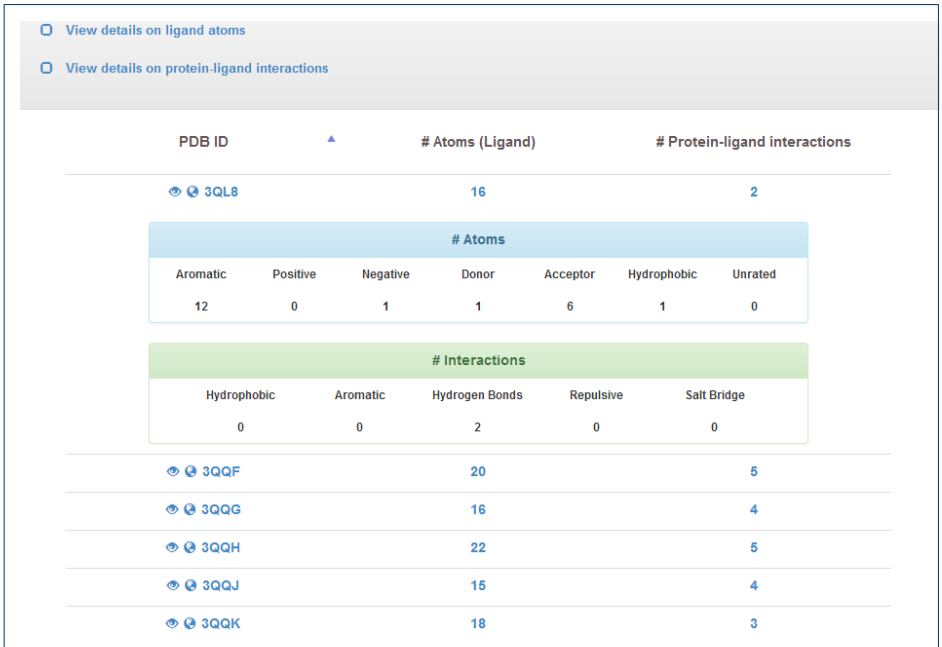


Figure 1 Dataset summary screen. It is an interactive table with summary of the number of ligand atoms and the number of protein-ligand interactions for each ligand. We also show a sub table with the number of each type of atoms or interactions. Users can compare the number of contacts established by the different ligands with the common receptor. By clicking on the eye icon in each line of the table and get a GLMol plugi-in with the protein and the ligand. There is also a link to the PDB web site where users can find the original data about each PDB file.

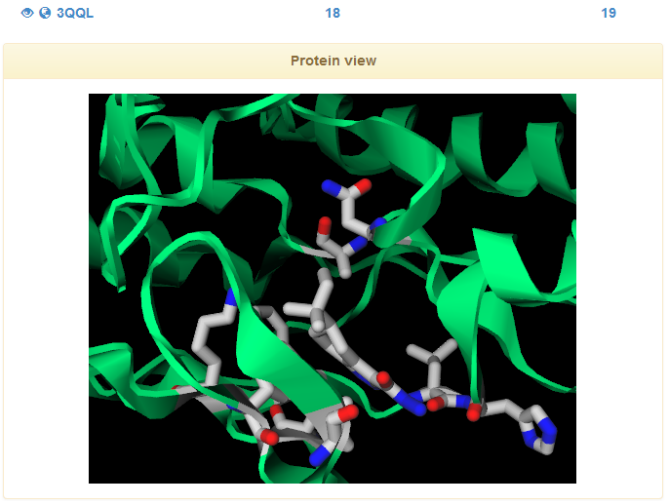


Figure 2 Protein and ligand structures visualization with GLMol plug-in. The protein is showed in cartoon and the ligand and the protein interacting residues in sticks colored in CPK. Thus users can interact with the structure by moving, rotating, and zooming the binding site. The set of PDB files were all superposed in a way that the are all in the same orientation making easier users comparisons.

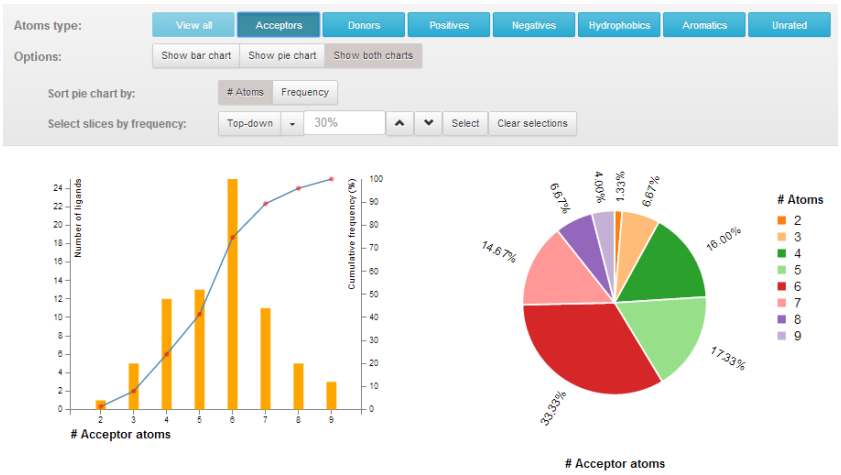


Figure 3 Graphical plot of atoms. Visual interactive interface where users can find the distribution of the number of ligands by the number of atoms of a specific type in a histogram. There is also a superposed line chart and a second y-axis representing the cumulative values of percentage frequencies. Notice that the chart is interactive and users can obtain details by passing the mouse over the bars or the points of the line. We also provided pie charts to depict the part-to-whole data relationship. They are also very interactive and users can not only obtain details on demand but also select a set of slices that sum up a certain percentage.

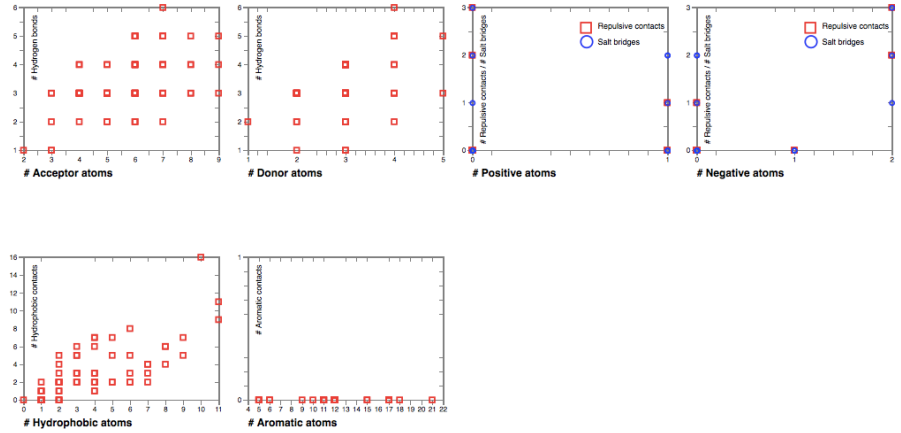


Figure 4 Correlation between the number of atoms of each type and the number of established interactions.

Atom	Type of interaction	# Ligands with which it interacts
LEU83		Total: 75 (100.00%)
N	Hydrogen bond	42
O	Hydrogen bond	75
GLU81		Total: 63 (84.00%)
O	Hydrogen bond	63
LEU134		Total: 44 (58.67%)
CD2	Hydrophobic	7
CD1	Hydrophobic	42

Figure 5 Interaction by residues screen. It is a very interactive table with color codes. The basic columns are Atom, Type of interaction and Number of interacting ligands. These lines can be grouped by header lines with the following columns: Residue and Percentage of interacting ligands. This grouping lines are colored according to a code that goes from blue (lower values) to orange (higher values). There are various sorting possibilities and the possibility of search in any of its fields.

Search: repulsive

Atom	Type of interaction	# Ligands with which it interacts
ASP86		Total: 24 (32.00%)
OD1	Repulsive	2
OD2	Repulsive	2
LYS89		Total: 10 (13.33%)
NZ	Repulsive	1
GLU8		Total: 2 (2.67%)
OE2	Repulsive	1
OE1	Repulsive	1

Figure 6 Another example of the Interaction by residues screen. Here we show an example where we searched by repulsive contacts only.

Search: repul

PDB id	CB	CD	CD1	CD2	CE	CE2	CG	CG1	CG2	CZ	N	NZ	O	OD1	OD2	OE1	OE2
3R7I (X9I)	ALA31 Hydrop. C7		LEU134 Hydrop. C7								LEU83 H. bond O3	LYS89 S. bridge O16	LEU83 H. bond N3 GLU81 H. bond N2			GLU8 Repuls. O17	GLU8 Repuls. O17 GLU8 Repuls. O16
3R83 (Z14)	ALA31 Hydrop. C15		LEU134 Hydrop. C15								LEU83 H. bond O26	LYS89 Repuls. N06	GLU81 H. bond N02 LEU83 H. bond N01				
3RK5 (07Z)					LYS89 Hydrop. C22							LYS89 S. bridge O23 LYS89 S. bridge O24	LEU83 H. bond N2 GLU81 H. bond N8	ASP86 Repuls. O24	ASP86 Repuls. O24		
3RM6 (18Z)	GLN131 Hydrop. C24 ALA31 Hydrop. C16	GLN131 Hydrop. C25 GLN131 Hydrop. C24	LEU134 Hydrop. C16		LYS129 Hydrop. C26						LEU83 H. bond O28	LYS129 Repuls. N08	GLU81 H. bond N01 LEU83 H. bond N03				

Figure 7 Interactions by ligands screen. It is a matrix where lines are the ligands and the columns are possible residue atoms. In the cells we present the residue in contact, the type of interaction and the ligand atom. It is a very simple way to show the data and to filter out unnecessary data.

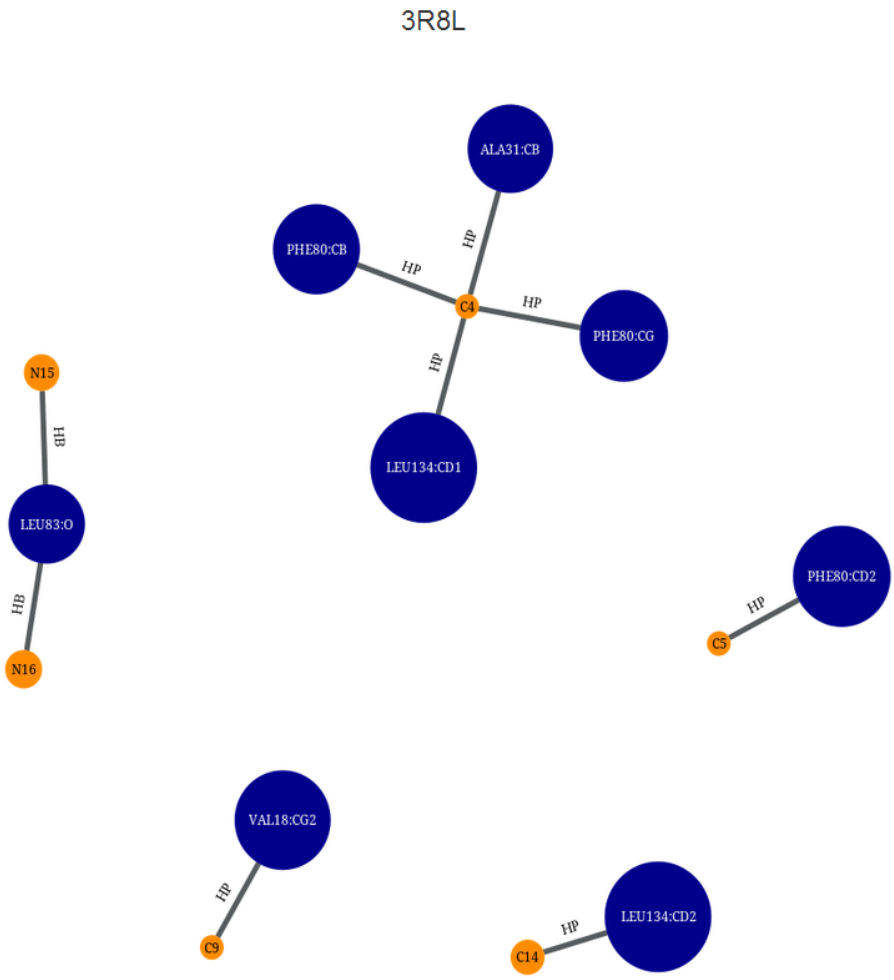
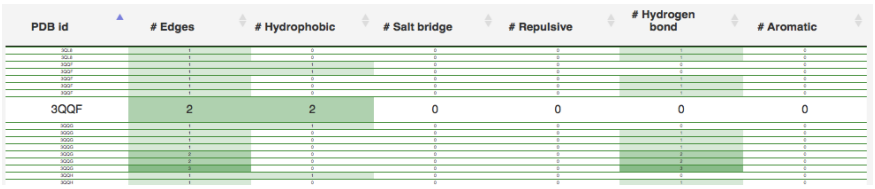


Figure 8 Protein-ligand contacts graph. In the Interactions by ligand screen, users can obtain a graph where blue nodes present protein residues name and number and atom and orange nodes depict ligand atoms.

# Edges	# Hydrophobic	# Salt bridge	# Repulsive	# Hydrogen bond	# Aromatic	Total in whole dataset
1	0	0	0	1	0	249
1	0	0	1	0	0	9
1	0	1	0	0	0	10
1	1	0	0	0	0	215
2	0	0	0	2	0	15
2	0	0	2	0	0	4
2	0	1	0	1	0	1
2	0	1	1	0	0	7
2	0	2	0	0	0	4
2	2	0	0	0	0	147
3	0	0	0	3	0	6
3	0	0	3	0	0	1
3	0	1	2	0	0	2
3	0	2	0	1	0	1
3	0	2	1	0	0	3
3	3	0	0	0	0	73
4	0	0	0	4	0	2
4	0	1	3	0	0	1
4	0	2	2	0	0	1
4	4	0	0	0	0	24
5	5	0	0	0	0	6

Figure 9 Protein-ligand subgraphs (summary). It is a heat map were users can see the summary of frequencies of each possible subgraph in the dataset in terms of the number of edges in general and each type: hydrophobic, salt bridge, repulsive, hydrogen bond, aromatic and the total frequency of occurrence in the dataset.



PDB id	# Edges	# Hydrophobic	# Salt bridge	# Repulsive	# Hydrogen bond	# Aromatic
3QQF	2	2	0	0	0	0

Figure 10 Protein-ligand subgraphs (all). It is a heat map were users can see, for each ligand (PDB file) the frequencies of each possible subgraph in the dataset in terms of the number of edges in general and each type: hydrophobic, salt bridge, repulsive, hydrogen bond and aromatic. As this heatmap has a high number of lines and does not fit in the screen, its cells are much smaller and we implemented a fish-eye based mouse over event as shown for the 3QQF line.

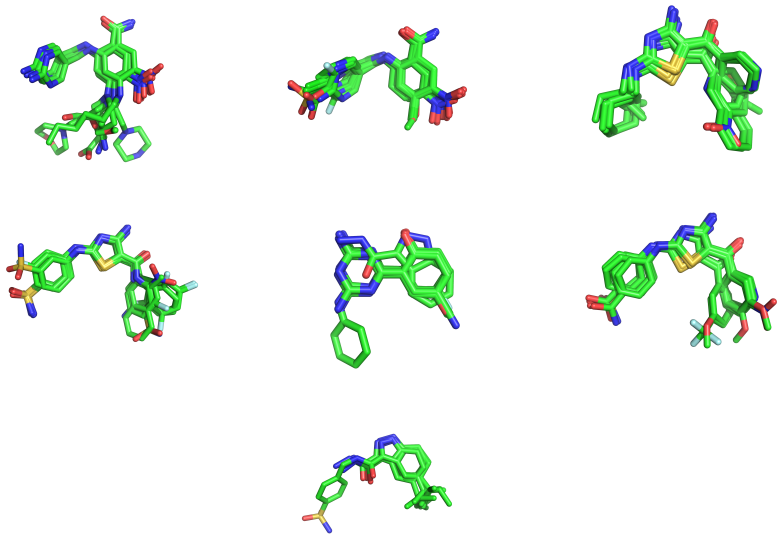


Figure 11 Strutural superposition of ligand clusters.

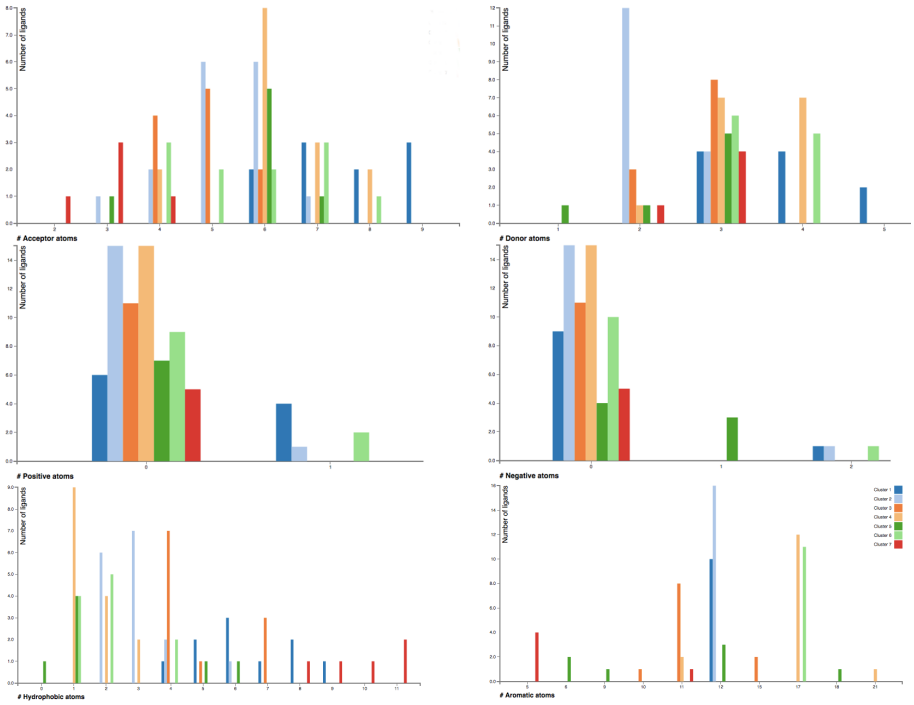


Figure 12 Distribution of types of atoms by clusters. They are distribution plots were users can compare the different cluster distribution.