CHARLES ABREU SANTANA

GREMLIN: UMA ESTRATÉGIA BASEADA EM MINERAÇÃO DE SUBGRAFOS PARA INFERIR PADRÕES DE INTERAÇÃO NA INTERFACE PROTEÍNA-LIGANTE

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae*

VIÇOSA
MINAS GERAIS – BRASIL
2017

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
S232g 2017

Santana, Charles Abreu, 1992-
Gremlin : uma estratégia baseada em mineração de subgrafos para inferir padrões de interação na interface proteína-ligante / Charles Abreu Santana. – Viçosa, MG, 2017.
xi, 92f. : il. (algumas color.) ; 29 cm.

Inclui apêndices.
Orientador: Sabrina de Azevedo Silveira.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f.49-54.

1. Mineração de dados (Computação). 2. Proteínas.
I. Universidade Federal de Viçosa. Departamento de Informática. Programa de Pós-graduação em Ciência da Computação.
II. Título.

CDD 22 ed. 006.312

CHARLES ABREU SANTANA

GREMLIN: UMA ESTRATÉGIA BASEADA EM MINERAÇÃO DE SUBGRAFOS PARA INFERIR PADRÕES DE INTERAÇÃO NA INTERFACE PROTEÍNA-LIGANTE

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae.*

APROVADA: 03 de março de 2017.

Fábio Ribeiro Cerqueira
Coorientador

Maria Cristina Baracat Pereira

Sabrina de Azevedo Silveira
Orientador

À família, aos amigos e a todos que dedicarem seu tempo à esta leitura.

"Tudo me é lícito, mas nem tudo me convém."
(Coríntios 6:12)

Agradecimentos

Agradeço primeiramente a Deus por permitir que eu desfrute desta passagem chamada vida. Como uma vez foi dito, "a vida é curta para ser pequena"(Benjamin Disraeli), logo viverei grandiosamente. Já basta que a vida seja curta para cogitar a possibilidade de apequená-la.

Agradeço à minha família pelo apoio, em especial à minha mãe Eliene, pela preocupação e atenção, aceitando, mesmo sob a dor da distância, que seu filho desbrave o mundo.

Agradeço à Universidade Federal de Viçosa e ao Departamento de Informática, junto aos seus professores, que sempre se dedicam ao máximo para passar o conhecimento adiante. Agradeço especialmente à professora Sabrina de Azevedo Silveira, pela paciência, inteligência e dedicação em guiar-me, de forma sábia, pelo árduo caminho acadêmico.

Agradeço aos amigos do mestrado, pelos bons momentos compartilhados com boa conversa, muita comida e alegria de sobra.

Agradeço ao financiamento fornecido pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), fundamental para que eu me dedicasse exclusivamente a este trabalho.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

SANTANA, Charles Abreu, M.Sc. Universidade Federal de Viçosa, março de 2017. **GREMLIN: Uma estratégia baseada em mineração de subgrafos para inferir padrões de interação na interface proteína-ligante**. Orientador: Sabrina de Azevedo Silveira. Coorientador: Fabio Ribeiro Cerqueira.

Interações proteína-ligante, de alta relevância em vários processos biológicos, são responsáveis pelo reconhecimento molecular, influenciando diretamente em mudanças de conformação das estruturas e, consequentemente, mudanças em sua atividade funcional. Entender essas interações é um passo importante para a predição de ligantes, identificação de alvos biológicos e projeto de fármacos. Esta dissertação propõe GReMLIN, uma estratégia baseada em mineração de subgrafos frequentes, para encontrar padrões em interações proteína-ligante. Aqui, investigamos se é possível encontrar padrões que caracterizam interações em um conjunto específico de proteínas. Se tais padrões existem, acreditamos que eles podem representar um passo importante na predição de interações. As interfaces proteína-ligante foram modeladas como grafos bipartidos, em que os vértices são átomos da proteína ou do ligante e as arestas são interações entre os átomos. Os vértices e arestas foram rotulados com suas propriedades físico-químicas. Um algoritmo de agrupamento foi executado sobre os dados dos grafos a fim de caracterizá-los de acordo com suas similaridades e diferenças e, em sequência, foi utilizado um algoritmo de mineração de subgrafos para buscar padrões relevantes nas estruturas de cada grupo. Para validar esta estratégia e verificar sua aplicabilidade em cenário real, foram coletados dados estruturais de complexos de proteínas com ligantes no Protein Data Bank. Foram usadas duas bases de dados, Ricina e CDK2, ambas com relevância biológica. GReMLIN foi capaz de encontrar subestruturas frequentes nos dados de Ricina e CDK2, contendo resíduos importantes determinados experimentalmente.

Abstract

SANTANA, Charles Abreu, M.Sc. Universidade Federal de Viçosa, March, 2017. **GREMLIN: A subgraph mining strategy based to infer interaction patterns in protein-ligand interface.** Adviser: Sabrina de Azevedo Silveira. Coadviser: Fabio Ribeiro Cerqueira.

Interaction between proteins and ligands are relevant in many biological process. Such interactions have gained more attention as the comprehension of protein-ligand molecular recognition is an important step to ligand prediction, target identification and drug design. This work proposes GreMLIN, a strategy to search patterns in protein-ligand interactions based on frequent subgraph mining. Here, we investigated if it is possible to find patterns that characterize protein-ligand interactions in a set of selected proteins. Moreover, if such patterns exist, we believe that they can represent an important step in the prediction of protein-ligand interactions. Our strategy models protein-ligand interfaces as bipartite graphs where nodes represent protein or ligand atoms, and edges represent interactions among them. Nodes and edges are labeled with physicochemical properties of atoms and a distance criteria. A clustering analysis is performed on graphs to characterize them according their similarities and differences, and a subgraph mining algorithm is applied to search for relevant patterns in protein-ligand interfaces in each cluster. We collected structural data of protein-ligand complexes in Protein Data Bank (PDB) to validate our strategy and show their applicability. Both datasets have biological relevance, but with different characteristics. Our strategy was able to find frequent substructures with considerable cardinality in the protein-ligand interfaces for the CDK and Ricin datasets.

Capítulo 1

INTRODUÇÃO

A bioinformática lida com o desenvolvimento e a aplicação de algoritmos e métodos de propósito geral para manipular a informação sobre dados biológicos, aplicando estes métodos no cenário real e criando novos conhecimentos [Xiong, 2006]. A bioinformática estrutural, ramo da bioinformática, trabalha com representação, armazenamento, recuperação, análise e visualização de informações estruturais tridimensionais de biomoléculas em escala atômica [Gu & Bourne, 2009].

A forma como as biomoléculas se organizam no espaço tridimensional é crucial para determinar suas funcionalidades [Nelson et al., 2014]. Quando moléculas interagem entre si, a sua configuração e conformação física são de máxima importância. Por exemplo, para um reagente ligar-se à sua respectiva enzima, ou um antígeno ligar-se a um anticorpo específico, é necessário que exista um reconhecimento molecular, isto é, as estruturas devem se encaixar com perfeita coesão. Portanto, estudar a estrutura tridimensional das biomoléculas é um passo importante para entender suas funcionalidades e interações.

O crescimento do número de estruturas macromoleculares tridimensionais (3D) disponíveis nas bases de dados e acessíveis na internet tem catalizado a evolução da bioinformática estrutural. Um exemplo é o Protein Data Bank (PDB) [1] [Berman et al., 2000], uma base de dados biomoleculares disponibilizada publicamente e com mais de 120.000 entradas de dados estruturais. À medida que a dimensão dos dados cresce, torna-se mais desafiador extrair informações dos mesmos. Assim, métodos que lidam com grandes volumes de dados tornam-se necessários, de forma que a informação seja disponibilizada de maneira inteligível para os pesquisadores de interesse.

A bioinformática estrutural vem experimentando avanços significativos com

[1] http://www.rcsb.org/pdb/statistics/holdings.do

pesquisas de natureza altamente interdisciplinar [Wei et al., 2015]. Consequentemente, isso vem despertando interesse de vários pesquisadores e gerando trabalhos promissores que resultam no progresso de métodos e aplicações. A grande promessa é que, mediante informação estrutural de alta resolução das biomoléculas, seja possível compreender e elucidar as funcionalidades dos sistemas biológicos [Gu & Bourne, 2009].

Apesar dos avanços, alguns desafios ainda permeiam o universo da bioinformática estrutural, como citados por Gu & Bourne [2009]. Primeiramente, a heterogeneidade em que os dados estruturais são apresentados, fazem com que os algoritmos, para seu processamento, sejam custosos, exigindo aproximações sem perda de informação importante. O espaço de busca para os problemas estruturais é amplo, pois suas variáveis são de natureza contínua, geralmente representadas por coordenadas Cartesianas. Além disso, dados estruturais demandam ferramentas de visualização para serem melhor compreendidos; a representação visual dos dados é de alto grau de importância no processo de análise. E finalmente, dados estruturais, como qualquer outro tipo de dado, podem conter ruídos e imperfeições que dificultam seu processamento.

A bioinformática estrutural lida com a predição e análise de estruturas 3D de macromoléculas como DNA, RNA e proteínas. Esta dissertação trata do universo proteico, mais precisamente das interações entre proteínas e ligantes. Nesse contexto, ligantes são pequenas moléculas que interagem com as proteínas através de ligações intermoleculares.

As proteínas desempenham papel fundamental nos processos biológicos, participando de atividades catalizadoras de reações químicas ou até mesmo na composição de estruturas supramoleculares das células [Williams & Daviter, 2013]. Compostos menores, denominados ligantes, conectam-se a uma proteína-alvo por meio de ligações não covalentes, ocasionando mudanças conformacionais na mesma e induzindo alterações em sua atividade [Gonçalves-Almeida et al., 2012]. As interações que ocorrem entre a proteína e o ligante são dirigidas por propriedades físico-químicas e estruturais que carregam características importantes, podendo auxiliar na elucidação de particularidades funcionais das biomoléculas envolvidas [Pires et al., 2013].

Com a riqueza de dados fornecidos pelas bases publicamente disponíveis como PubChem [Kim et al., 2015], Zinc [Sterling & Irwin, 2015], DrugBank [Wishart et al., 2006] e PDB [Berman et al., 2000], podemos tentar inferir sobre como os ligantes interagem com sua proteína-alvo. Uma caracterização detalhada das interações proteína-ligante pode auxiliar na compreensão do reconhecimento molecular,

entender as funcionalidades proteicas e ajudar na predição de novos ligantes para proteínas [Salentin et al., 2015].

A predição de ligantes para uma proteína é uma das tarefas mais desafiadoras na bioquímica, com implicação direta na química farmacêutica, na descoberta de funções proteicas, na identificação de alvos biológicos e no projeto de fármacos [Konc & Janežič, 2014]. Um possível caminho para caracterizar interações e posteriormente predizê-las é através da localização de suas estruturas conservadas. A descoberta de padrões conservados é importante para entender como uma biomolécula alvo está ligada a um composto qualquer [Gonçalves-Almeida et al., 2012]. Um padrão pode ser considerado como um conjunto conservado de atributos da interface de ligação usados para explicar e entender a mesma [Gonçalves-Almeida et al., 2012].

Observando essa perspectiva dos padrões de interação, este trabalho propõe uma estratégia baseada em mineração de grafos, denominada GReMLIN, capaz de inferir padrões existentes na interface de interação entre a proteína e seu respectivo ligante. Grafos são estruturas capazes de modelar objetos e seus relacionamentos (como átomos e interações), e é uma ferramenta poderosa, bem estruturada, contendo teoremas e algoritmos capazes de resolver diversos problemas [Cormen et al., 2010].

Ao modelar a interface de ligação entre o ligante e sua respectiva proteína como grafos, será possível usar algoritmos da literatura, como a mineração de subgrafo frequente (MSF), para extrair e manipular estruturas relevantes (padrões de interação) dos dados. A MFS consegue extrair subestruturas frequentes de dados modelados como grafos. Acreditamos que com esses padrões, caso existam, será possível entender e identificar particularidades que contribuam significantemente para a interação proteína-ligante. Além disso, como trabalho futuro, a devida caracterização destas estruturas relevantes pode auxiliar na tentativa de predizer potenciais ligantes para uma proteína alvo.

1.1 Objetivos

O objetivo deste trabalho foi projetar, implementar e avaliar uma estratégia baseada em mineração de subgrafo frequente para inferir padrões de interação na interface de complexos entre proteína e ligante.

Os objetivos específicos foram:

- Implementar um conjunto de programas para modelar os arquivos estruturais em grafos de interações proteína-ligante, rotulando-os adequadamente com

suas respectivas propriedades físico-químicas.

- Sumarizar a base de grafos construída usando uma matriz de atributos e usar métodos de redução de dimensionalidade e eliminação de ruídos na matriz;
- Verificar se a matriz que sumariza os dados pode ser utilizada para realizar o agrupamento dos grafos.
- Usar a matriz que sumariza os grafos para particioná-los;
- Executar o algoritmo de mineração de subgrafo frequente, em cada grupo separadamente, para extrair padrões dos grafos;
- Filtrar os grafos maximais dentre os padrões obtidos;
- Mapear os padrões resultantes para os grafos de entrada, identificando onde estes se encontram nas estruturas do PDB.
- Propor e implementar estratégia de visualização dos dados e resultados.

1.2 Organização do trabalho

Essa dissertação foi organizada da seguinte forma: no Capítulo 2, *Referencial Teórico*, serão explicados fundamentos teóricos necessários para o entendimento da estratégia proposta, assim como trabalhos da literatura relacionados ao GReMLIN. O Capítulo 3, *Métodos*, descreve os passos para se alcançar as estruturas conservadas nas interações, explicando como cada técnica foi utilizada durante o fluxo da estratégia. Os resultados de experimentos realizados são descritos no Capítulo 4. No Capítulo 5, são feitas algumas considerações e descrições sobre as próximas etapas do trabalho.

Capítulo 2

REFERENCIAL TEÓRICO

Neste capítulo, serão descritos os fundamentos necessários para compreender o GReMLIN. Algumas definições importantes serão apresentadas, assim como algoritmos e estratégias utilizadas, desde a modelagem das biomoléculas até a obtenção dos padrões.

2.1 A proteína e sua estrutura

As proteínas são as macromoléculas biológicas mais abundantes, ocorrendo em todas as células, e constituído maior fração celular (além da água). Elas desempenham diversas funções importantes, atuando em atividades catalizadoras de reações químicas, constituindo estruturas supramoleculares, atuando como receptoras de sinal, ou transportando substâncias específicas para dentro ou para fora da célula [Williams & Daviter, 2013]. As proteínas se apresentam em forma de fibras musculares, anticorpos, hormônios, antibióticos e venenos. Também pode ser encontrada na composição de estruturas encontradas no cotidiano, como na lente dos olhos, penas, teias de aranha, chifres, unhas e cabelo [Nelson et al., 2014].

Proteínas são polímeros formados por partes monoméricas denominadas aminoácidos. A estrutura proteica é composta por resíduos de aminoácidos unidos em sequência linear característica através de ligações covalentes. Os aminoácidos são compostos químicos contendo átomos de carbono (C), hidrogênio (H), oxigênio (O), nitrogênio (N) e algumas ocorrências de enxofre (S). A estrutura de um aminoácido, mostrado na Figura 2.1, é caracterizada pela presença do grupo amina (NH_2) e do grupo carboxílico (COOH), ambos ligados a um mesmo átomo de carbono denominado de carbono α. Este carbono também é ligado a um átomo de hidrogênio e a

uma cadeia lateral, que é representada pela letra R. Os aminoácidos diferem uns dos outros em suas cadeias laterais (grupos R) [Buxbaum, 2015].

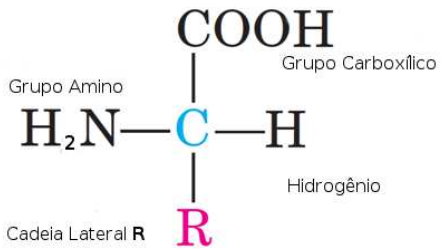


Figura 2.1. Estrutura geral de um aminoácido.

Os aminoácidos, quando unidos através de ligações peptídicas, são chamados de *resíduos de aminoácidos*, pois durante a ligação há perda de elementos de água dos compostos originais [Nelson et al., 2014]. Uma sequência de aminoácidos ligados por interações peptídicas forma um polipeptídeo (Figura 2.2). Uma proteína pode conter uma ou mais cadeias polipeptídicas.

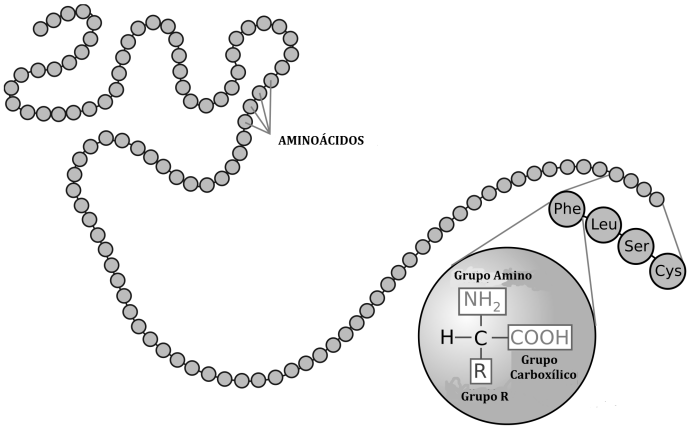


Figura 2.2. Esquema de uma cadeia conectada de aminoácidos.

Todas as proteínas são compostas por apenas 20 tipos diferentes de aminoácidos [Fromm & Hargrove, 2012]. Cada um desses aminoácidos comuns, são diferenciados unicamente pela cadeia lateral R. Por motivo de simplificação foram atribuídos aos aminoácidos abreviações de três letras, e representação usando somente um caractere, com o objetivo de reduzir o uso de memória em procedimentos de bioinformática. Na Tabela 2.1, está listado os nomes dos 20 aminoácidos comuns e suas respectivas abreviações de três caracteres e um caractere respectivamente. Todas as estruturas dos aminoácidos comuns em proteínas estão representados na Figura 2.3.

Tabela 2.1. Lista de aminoácidos e suas respectivas abreviações.

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

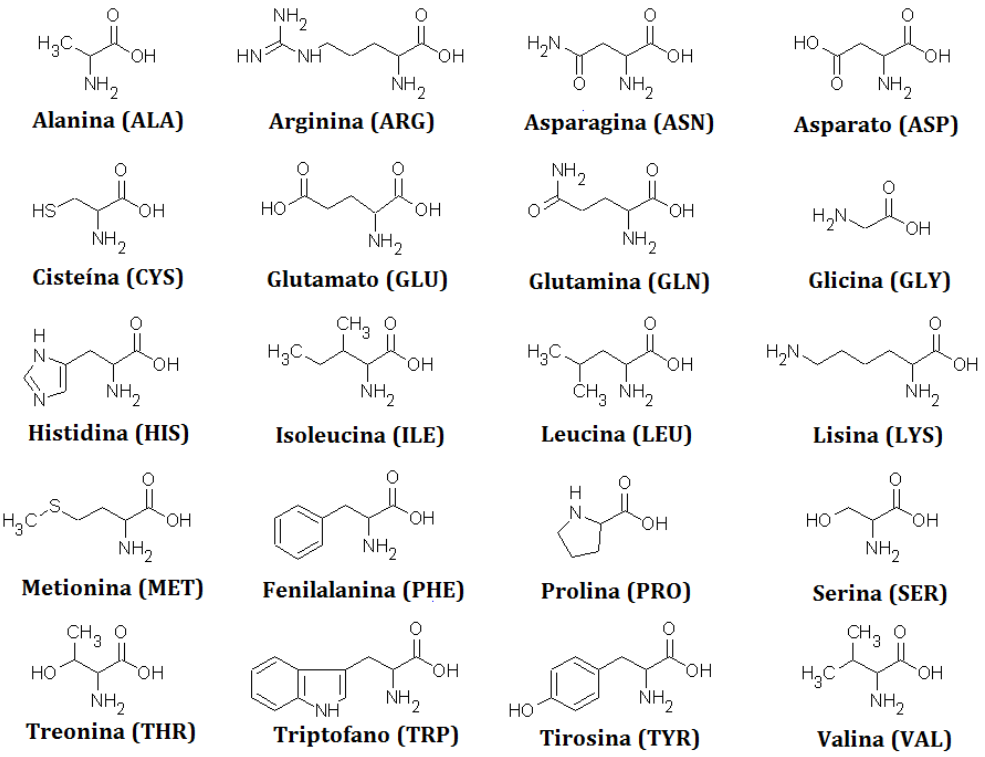


Figura 2.3. Os 20 aminoácidos comuns de proteínas.

A estrutura de uma proteína pode ser observada em quatro níveis distintos (Figura 2.4). Na estrutura primária a proteína é vista em termo de sua sequência de resíduos de aminoácidos em uma cadeia polipeptídica. A estrutura secundária visualiza a proteína em termos de arranjos estruturais dos resíduos, como hélices, folhas e alças, que dão origem a padrões estruturais estáveis. A estrutura terciária descreve todo o arranjo tridimensional da conformação dos átomos da proteína,

descrevendo todos o aspectos do enovelamento tridimensional do polipeptídeo. E finalmente, a estrutura quaternária, que observa a proteína em termos de suas subunidades (cadeias polipeptídicas).

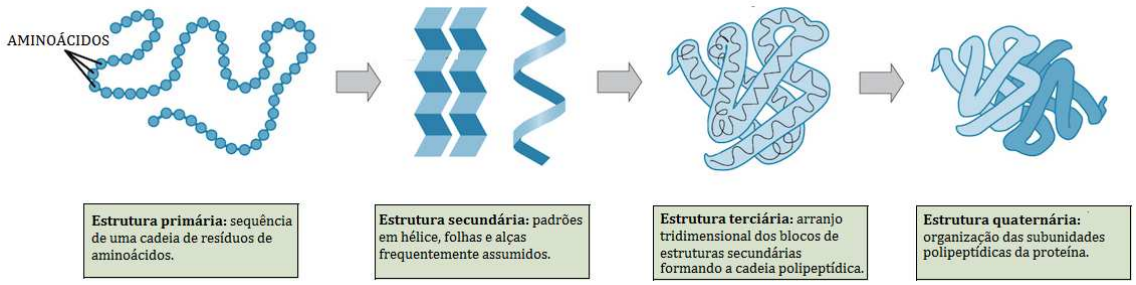


Figura 2.4. Níveis de estrutura nas proteínas.

As proteínas diferem-se pelo número e sequência de seus resíduos de aminoácidos. Cada proteína possui uma sequência de aminoácidos específica que confere uma determinada estrutura tridimensional, e essa estrutura, por sua vez, confere uma função específica à proteína [Fromm & Hargrove, 2012]. Consequentemente, pode-se inferir que proteínas com funções diferentes possuem sequencias de aminoácidos diferentes. Uma família de proteínas é constituída por um grupo de proteínas que compartilham uma semelhança em sua estrutura primária, funcionalidade ou características estruturais [Nelson et al., 2014].

A estrutura proteica é dinâmica, sendo que suas funções dependem de interações com outras moléculas, e essas interações induzem mudanças importantes na conformação da proteína [Williams & Daviter, 2013]. Essas moléculas que interagem com a proteína são chamadas de ligantes. Um ligante pode ser qualquer tipo de molécula, inclusive outra proteína. Contudo, para este trabalho, consideramos ligantes como pequenas moléculas, não proteicas, que interage em uma conformação energeticamente favorável com o sítio de ligação de proteínas. O sítio de ligação é uma região da proteína que é complementar ao ligante em tamanho, forma, carga e caráter hidrofílico (afinidade de interação com a água) e hidrofóbico (não possui afinidade com a água) [Buxbaum, 2015]. A interação que ocorre no sítio de ligação é totalmente específica. Além disso, pode existir vários sítios separados em uma proteína para vários ligantes diferentes.

2.2 Interações proteína-ligante

Interações proteína-ligante desempenham papel fundamental em vários sistemas biológicos, sendo responsáveis por processos como a transmissão de sinais, através da complementaridade molecular, e mudanças conformacionais, induzidas por ligações não covalentes no sítio de ligação [Williams & Daviter, 2013]. Essas mudanças conformacionais modificam as funcionalidades proteicas, ativando-as ou inibindo-as.

Entender interações que ocorrem entre o ligante e seu receptor é a base para a química medicinal [Mannhold et al., 2006]. A descoberta de novos medicamentos para tratar doenças importantes é um dos maiores desafios na pesquisa farmacêutica e, segundo Mannhold et al. [2006], estudos baseados na estrutura molecular vêm ganhando destaque, sendo que vários fármacos comercializados estão relacionados ao sucesso desta abordagem estrutural.

Indústrias farmacêuticas e grupos de pesquisa interessados em descobrir novos candidatos a medicamentos estão procurando métodos mais rápidos, efetivos e confiáveis [Guedes et al., 2014]. Uma das estratégias emergentes, com grande importância na busca de compostos candidatos, é a triagem virtual [Cheng et al., 2012]. A triagem virtual é usada para analisar computacionalmente um número grande de compostos e selecionar os mais promissores, em termos de afinidade com um alvo biológico em particular, levando em consideração alguns critérios pré-definidos. Um esquema de funcionamento desta técnica está representado na Figura 2.5. A seleção é baseada em uma função de avaliação usada para mensurar a afinidade entre o composto candidato e o alvo biológico [Huang et al., 2010].

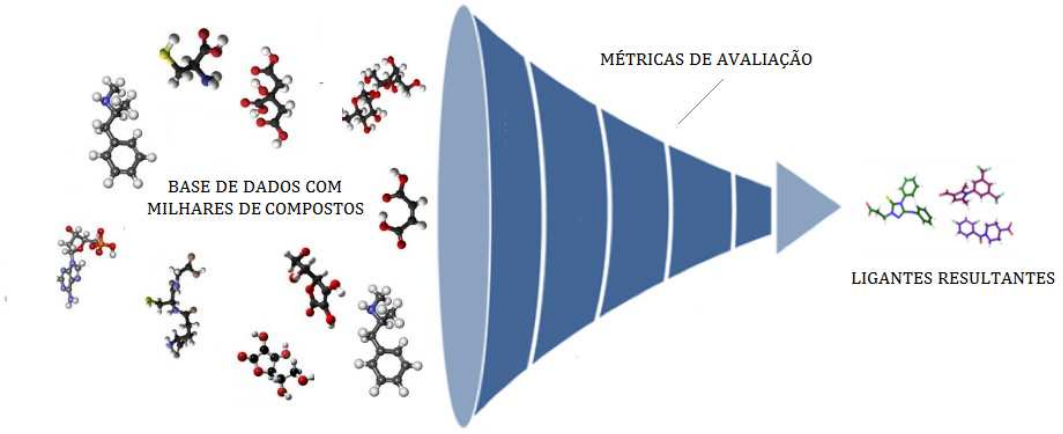


Figura 2.5. Esquema simplificado de uma técnica de triagem virtual.

A busca por ligantes na triagem virtual pode ser feita usando diversas bases de dados que estão disponíveis online, como: PubChem [Kim et al., 2015], Zinc [Sterling

& Irwin, 2015] e DrugBank [Wishart et al., 2006]. Os compostos selecionados são sujeitos a etapas experimentais como síntese química e estudos *in vitro* e *in vivo* [Guedes et al., 2014]. Baseando-se nesses resultados, ligantes mais promissores serão potenciais candidatos a novos medicamentos.

A compreensão das interações proteína-ligante pode apoiar na predição de potenciais ligantes para alvos de interesse, e também apoiar no processo de descoberta de fármacos. Nesse contexto, encontrar meios que possam caracterizar tais interações pode elucidar particularidades responsáveis pelo reconhecimento estrutural entre as biomoléculas. Desse modo, o trabalho descrito nessa dissertação está empenhado em investigar se existem padrões de interação na interface proteína-ligante em nível atômico e, se tais padrões existirem, como eles podem influenciar na interação entre a proteína-alvo e o ligante. Para isso, será utilizada uma modelagem baseada em grafos e um processo de extração de padrões estruturais baseados em técnicas de mineração de dados descritos nas próximas seções.

2.3 Grafos

Uma maneira simples e robusta de se representar objetos e seus relacionamentos é através dos grafos [Sedgewick & Flajolet, 2013]. Grafo é uma estrutura de dados poderosa e extensivamente usada no estudo de redes de transporte e comunicação, circuitos elétricos, componentes químicos e até mesmo para modelar dados estruturais de proteínas [Vishveshwara et al., 2002]. Algoritmos e métodos de mineração baseados em grafos tornaram-se úteis para pesquisas em sistemas biológicos como, por exemplo, a modelagem e investigação de estruturas tridimensionais de macromoléculas [Saidi et al., 2009; Shen & Guda, 2014].

Um grafo é definido como um par $G = (V, E)$, onde V é um conjunto vértices e E um conjunto de arestas que representam uma relação binária em V [Cormen et al., 2010]. Dois vértices conectados por uma aresta são denominados adjacentes.

Grafos podem ser dirigidos ou não-dirigidos. Em um grafo dirigido, ou digrafo, cada aresta é um par ordenado de vértices. Se (u, v) é uma aresta de um digrafo $G = (V, E)$, dizemos que (u, v) deixa u e incide em v. A Figura 2.6 (a) é uma representação gráfica de um digrafo $V = \{1, 2, 3, 4\}$, $E = \{(1, 2), (2, 4), (4, 2), (3, 4), (3, 1)\}$. Os vértices são representados por círculos, enquanto as arestas são representadas por arcos.

Um grafo G é dito não-dirigido caso não possua uma relação de ordem entre o par de vértices de uma aresta qualquer. Desse modo, uma aresta é um conjunto

$\{u, v\}$, onde $u, v \in V$ e $u \neq v$. Por convenção é usada a notação (u, v) para representar $\{u, v\}$, em que (u, v) e (v, u) são consideradas como a mesma aresta [Cormen et al., 2010; Sedgewick & Flajolet, 2013]. A Figura 2.6 (b) mostra a representação gráfica de um grafo não-direcionado $V = \{1, 2, 3, 4\}$, $E = \{(1, 2), (1, 3), (2, 4), (3, 4)\}$

Um grafo rotulado contém rótulos associados aos seus vértices e arestas. Uma função de rotulagem L é usada para associar rótulos aos elementos. Usa-se $L(u)$ para denotar o rótulo de um vértice $u \in V$ e $L(u, v)$ para denotar o rótulo de uma aresta $(u, v) \in E$ [Zaki et al., 2014]. Na Figura 2.6 (c), por exemplo, é mostrado um grafo rotulado $V = \{1, 2, 3, 4\}$, $E = \{(1, 2), (1, 4), (2, 4), (3, 4)\}$ e seus rótulos, $L(1) = A, L(2) = B, L(3) = B, L(4) = C, L(1, 2) = x, L(1, 4) = y, L(2, 4) = y, L(3, 4) = x$.

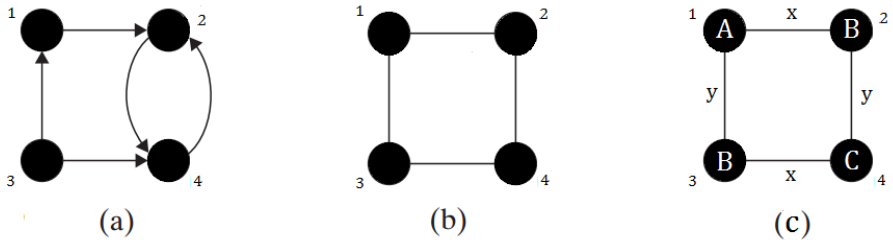


Figura 2.6. Representação de um digrafo (a), um grafo não-direcionado (b) e um grafo rotulado (c).

Segundo [Elseidy et al., 2014], um grafo qualquer $S = (V_S, E_S, L_S)$ é dito *subgrafo* de um grafo $G = (V, E, L)$ se e somente se satisfaz às seguintes regras:

1. $V_S \subseteq V$;
2. $E_S \subseteq E$;
3. $\forall\ \mu \in V_S \cup E_S,\ L_S(\mu) = L(\mu)$.

Em outras palavras, subgrafos são fragmentos de um grafo. A regra 1 diz que para um subgrafo S pertencer a um grafo G, o conjunto de vértices de S deve estar contido no conjunto de vértices de G. O mesmo vale para o conjunto de arestas na regra 2. A regra 3 diz que para todos os elementos de S (vértices e arestas), a rotulagem deve combinar de forma exata com os elementos de G. A Figura 2.7 mostra o exemplo de um grafo e alguns de seus subgrafos.

Um *caminho* entre dois vértices u e v de um grafo G é definido como uma sequência de vértices adjacentes que começa em u e termina em v. Quando todos os vértices são distintos em um caminho, este é chamado de simples. Um grafo G é denominado *conexo* se para todo par de vértices u e v de G existe um caminho de

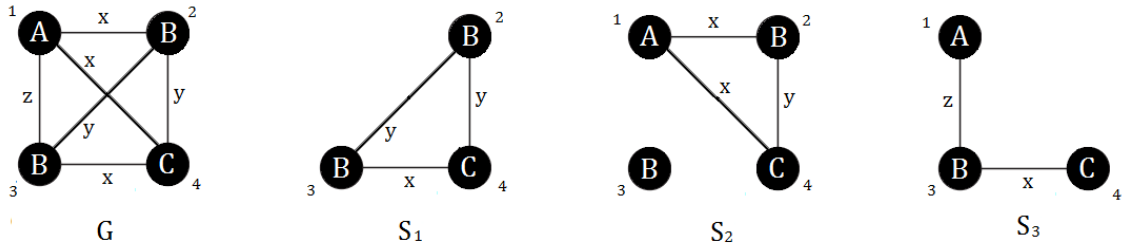


Figura 2.7. Representação de um grafo G e seus subgrafos S_1, S_2 e S_3

u para v. Por exemplo, na Figura 2.7, a sequência de vértices $(1, 2, 3)$ formam um caminho em G. Também, G é considerado um grafo conexo, pois para cada par de seus vértices existe um caminho. Porém, a conectividade não ocorre no subgrafo S_2 pois, partindo dos vértices 1, 2 ou 4, não é possível formar um caminho até o vértice 3, logo S_2 é desconexo.

Se um grafo não é conexo, suas partes são denominadas componentes conexas. O conjunto de componentes conexas de um grafo desconexo é formado pelos seus subgrafos maximais conectados. Assim, no grafo S_2 da Figura 2.7, existem duas componentes conexas, uma representada pelo subgrafo formado pelos vértices 1 e 3 e a outra pelo subgrafo formado pelos vértices 2 e 4.

No contexto biológico, a proteína pode ser enxergada como um conjunto de elementos relacionados (aminoácidos, átomos) [Saidi et al., 2009]. Portanto, ela pode ser facilmente transformada em um grafo, onde os vértices podem ser fragmentos da proteína, como átomos, por exemplo, e arestas podem representar interações entre estes fragmentos.

Ligantes similares possuem sítios de ligação parecidos em relação à sua forma e propriedades físico-químicas [Pires et al., 2013]. Isso sugere que ao caracterizar tais similaridades, é possível obter conhecimento que auxilie em tarefas maiores como na predição de ligantes para proteínas, identificação de alvos biológicos e descoberta de medicamentos.

A modelagem de complexos de proteínas com ligantes através de grafos permite explorar técnicas da literatura para encontrar padrões conservados nestas estruturas. Pela simplicidade que os grafos conseguem representar relacionamentos complexos entre objetos em diversas aplicações, encontrar estruturas relevantes em grafos tornou-se de muito interesse para pesquisadores [Shen & Guda, 2014].

Muitos algoritmos que extraem estruturas relevantes em base de grafos têm sido implementados, como: FFSM [Huan et al., 2003], gSpan [Yan & Han, 2002a], SPIN [Huan et al., 2004], MARGIN [Thomas et al., 2010] e Grami [Elseidy et al.,

2014], cada um atendendo a uma característica específica do problema de mineração em grafos, o que será melhor explanado na próxima seção.

2.3.1 Mineração de subgrafo frequente

Uma das técnicas usadas para encontrar padrões relevantes em grafos é denominada como *mineração de subgrafo frequente* [Jiang et al., 2013]. Essa é uma área de estudo que tem como função encontrar subestruturas que apareçam frequentemente em uma base de grafos. Por ser frequente, acredita-se que tais subestruturas possam representar características importantes dos dados.

Para encontrar subgrafos frequentes, é necessário algum mecanismo de comparação que possibilite verificar se dois grafos quaisquer são ou não iguais. Esta verificação é denominada de *teste de isomorfismo* e tem fundamental importância na mineração de subgrafos, pois é a partir de comparações sucessivas que podemos contabilizar a ocorrência de um dado subgrafo e dizer se o mesmo é frequente ou não [Jiang et al., 2013; Lee et al., 2012].

Dois grafos $G_1 = (V_1, E_1, L_1)$ e $G_2 = (V_2, E_2, L_2)$, são isomorfos se e somente se existe uma bijeção $f : V_1 \to V_2$ satisfazendo:

1. $(u, v) \in E_1 \iff (f(u), f(v)) \in E_2$
2. $\forall u \in V_1, L(u) = L(f(u))$;
3. $\forall (u, v) \in E_1, L(u, v) = L(f(u), f(v))$.

A bijeção f é dita um isomorfismo entre G_1 e G_2. Em outras palavras, o isomorfismo f preserva as adjacências entre as arestas assim como os rótulos dos vértices e arestas [Zaki et al., 2014].

Um *isomorfismo de subgrafo* é quando desejamos saber se um subgrafo G_1 está contido completamente em um grafo G_2 [Jiang et al., 2013]. Um isomorfismo de subgrafo de G_1 em G_2 ocorre quando existe um subgrafo $\mathbf{g} \subseteq G_2$ tal que G_1 é isomorfo a $\mathbf{g}$ [Huan et al., 2003]. Na Figura 2.8 é exibido um exemplo de isomorfismo de subgrafo onde o grafo (b) é subgrafo isomorfo de (a), enquanto o grafo (c) não é, pois sua estrutura topológica (adjacência destacada) não é preservada em relação ao grafo (a). Isomorfismo de subgrafos é considerado como um problema NP-difícil, em outras palavras, é um processo muito custoso computacionalmente para instâncias de grande magnitude [Yan & Han, 2002a].

Existem duas abordagens no processo de mineração de subgrafo frequente, a *abordagem transacional* e de *grafo único* [Jiang et al., 2013]. Na forma transacional,

os dados de entrada são vários grafos denominados transações. Para o formato de grafo único, como o nome sugere, o dado de entrada é um único grafo. Para este trabalho, será considerada a abordagem transacional, pois os dados de entrada para a realização da estratégia são compostos por vários grafos representando as interações na interface proteína-ligante.

Em uma base de grafos transacional, um subgrafo é considerado frequente quando seu percentual de ocorrências na base de grafos for superior a um limite inferior denominado *limiar de suporte*. Minerar subgrafo frequente é encontrar todos os subgrafos que aparecem em uma base de dados conforme um limiar de suporte pré-definido denominado **minSup**. Seja uma base de dados $D = \{G_0, G_1, ..., G_n\}$ e um subgrafo g, o número de grafos (transações) contidos em D, onde g é um subgrafo, é representado como $\sigma(g)$. No geral, o problema de mineração de subgrafo frequente é encontrar qualquer subgrafo g em que $\frac{\sigma(g)}{|D|} \geq$ **minSup**.

Na Figura 2.9, no quadro 1, tem-se uma base de grafos contendo 3 grafos (G1, G2 e G3). Considere também que as arestas possuem o mesmo rótulo, a fim de simplificação. No quadro 2 encontram-se os subgrafos frequentes da base transacional do quadro 1 com um suporte = 0.6, ou seja, subestruturas que aparecem nos dados com uma ocorrência maior ou igual a 60%. Os padrões triviais, isto é, aqueles constituídos por um só vértice, foram omitidos.

À medida que o tamanho dos grafos ou da base de dados aumenta, o número de padrões frequentes cresce exponencialmente [Huan et al., 2004]. Uma forma de diminuir expressivamente o número de subgrafos resultantes da mineração, sem perda de informação, é através dos subgrafos frequentes maximais. Um grafo g é dito frequente maximal se satisfaz as seguintes condições:

1. g deve ser frequente;

2. não deve existir nenhum grafo frequente do qual g seja um subgrafo;

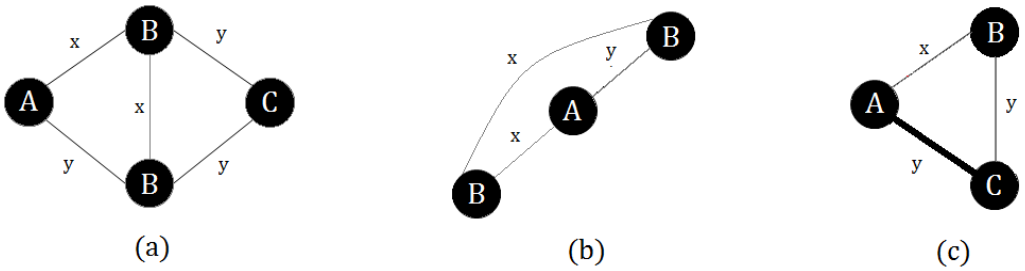


Figura 2.8. Exemplo de isomorfismo de subgrafo.

Segundo [Koyutürk et al., 2004], em redes biológicas grafos maximais são mais interessantes para análise. Os resultados de um algoritmo de mineração de subgrafo frequente são repetitivos em suas estruturas resultantes. Muitos dos padrões contidos nos resultados estão embutidos, ou seja, são subgrafos de padrões frequentes maiores. Extrair somente os subgrafos maximais pode reduzir significantemente a quantidade de resultados e, além disso, sem perda de informação, já que padrões menores podem ser obtidos a partir dos maximais. Tomando como exemplo a Figura 2.9, se extraídos os padrões maximais, o único padrão de saída seria o grafo **g1**. Todos os demais subgrafos frequentes estão embutidos em **g1**.

Existem vários trabalhos na literatura que contemplam a mineração de subgrafos. Um dos mais citados segundo Jiang et al. [2013] é o gSpan [Yan & Han, 2002a], que através de uma estratégia de busca em profundidade consegue representar os grafos de uma forma que deixa a extração de padrões mais eficiente. Para atender às necessidades da estratégia proposta nesta dissertação, utilizamos o algoritmo gSpan.

Outras abordagens como SPIN [Huan et al., 2004] e MARGIN [Thomas et al., 2010], tentam diminuir o número de dados na saída do algoritmo, extraindo somente grafos maximais, para melhorar a qualidade dos resultados. Alguns algoritmos de mineração de subgrafos, como o Subdue [Ketkar et al., 2005], incluem outras restrições além do suporte. Características como número de vértices e densidade, por exemplo, também podem ser usadas na extração de subestruturas. A estratégia proposta por Shelokar et al. [2013] utiliza heurísticas evolucionárias para diversificação das soluções, para que estas possam também atender a diferentes restrições.

Acreditamos que modelar estruturas biomoleculares como grafos e extrair suas

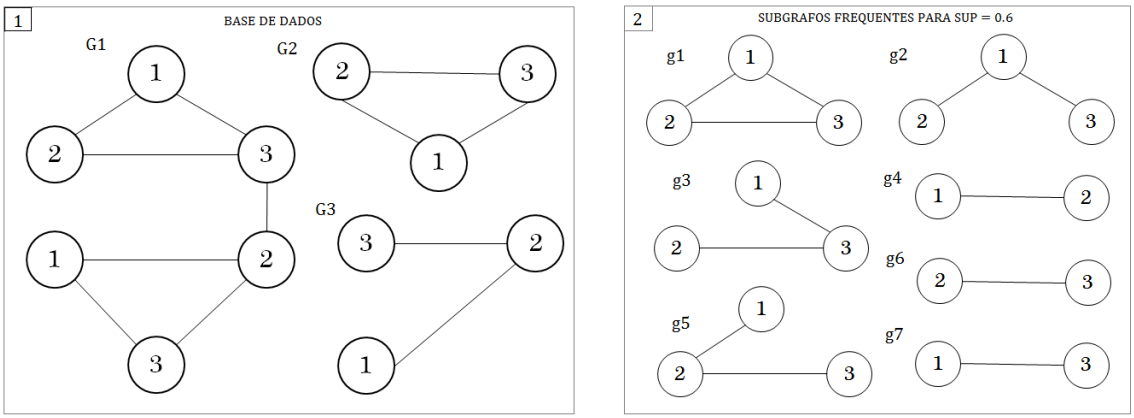


Figura 2.9. Base de grafos transacional (Quadro 1) e seus subgrafos frequentes para um suporte de 0.6, ou ocorrência maior ou igual a 60 % (Quadro 2).

subunidades conservadas aplicando algoritmos de mineração, mediante uma profunda análise, gerará novos conhecimentos, importantes na caracterização das interações proteína-ligante. Além disso, em trabalhos futuros, essas características inferidas podem auxiliar na busca por novos ligantes para proteínas de interesse.

2.4 Análise de agrupamentos

Organizar objetos que compartilham características semelhantes em grupos é uma estratégia muito utilizada para apoiar a compreensão destes objetos. Agrupar elementos para caracterizá-los é uma das formas como as pessoas descrevem o mundo [Rokach, 2009]. Análise de agrupamentos é frequentemente utilizada em diversas áreas como biologia, ciência da informação, segurança e reconhecimento de padrões em imagens [Han et al., 2011].

No contexto da mineração de dados, grupos são classes em potencial, e análise de agrupamento é o estudo das técnicas para encontrar classes automaticamente [Tan et al., 2006]. O objetivo destas técnicas é dividir os dados em grupos, de forma não supervisionada, observando somente as informações encontradas nos dados que descrevem os objetos e seus relacionamentos, para tentar capturar a estrutura natural dos mesmos.

Existem várias estratégias para agrupar dados. Contudo dependendo da natureza dos dados, alguns métodos podem se sobressair. No geral, a ideia é fazer com que objetos de um mesmo grupo sejam similares entre si e dissimilares a objetos de outros grupos. Porém, usar uma técnica de agrupamento não irá garantir o melhor particionamento dos dados, sendo que a definição de um grupo depende da natureza dos dados e resultados desejados [Tan et al., 2006]. Por exemplo, na Figura 2.10 é possível notar que, para um mesmo conjunto de dados, é viável criar grupos de maneiras diferentes e mesmo assim ser coerente na divisão.

Han et al. [2011] descreveram algumas características das técnicas fundamentais para agrupar objetos de dados:

- **Métodos baseados em protótipo:** em técnicas baseadas em protótipo, um grupo é um conjunto de objetos, em que cada objeto é mais similar ao protótipo que define seu respectivo grupo do que protótipos de outros grupos. O protótipo é um ponto representativo utilizado para definir um grupo. A Figura 2.11 mostra um exemplo de agrupamento baseado em protótipo. Cada objeto dos dados é atribuído a um grupo baseando-se na proximidade com o elemento representativo (protótipo).

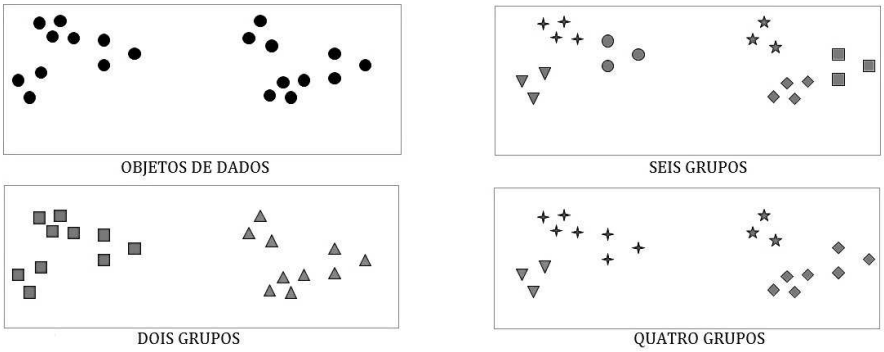


Figura 2.10. Diferentes agrupamentos para um mesmo conjunto de dados. (Modificado de [Tan et al., 2006]).

- **Métodos hierárquicos:** dados p pontos em um espaço n-dimensional, esta abordagem cria uma sequência de partições aninhadas que podem ser graficamente visualizadas através de uma árvore chamada *dendograma* [Zaki et al., 2014]. O dendograma exibe o relacionamento entre os clusters e sub-clusters na ordem em que eles são unidos (abordagem aglomerativa) ou particionados (abordagem divisionista) [Tan et al., 2006]. Um exemplo de dendograma e seu respectivo agrupamento aninhado é mostrado na Figura 2.12.

- **Métodos baseados em densidade:** esses métodos usam a densidade local para determinar os grupos. O objetivo é separar regiões de alta densidade das que possuem densidade inferior. Esse tipo de abordagem leva vantagem em ralação às técnicas que usam distância como métrica de similaridade, pois elas conseguem identificar grupos com formas diversificadas (Figura 2.13). Porém, em grupos com densidade heterogênea esse tipo de estratégia pode não ser eficaz.

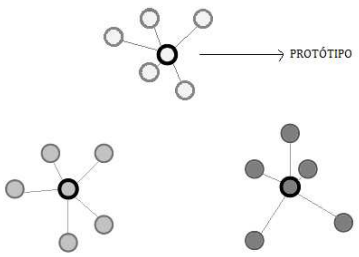
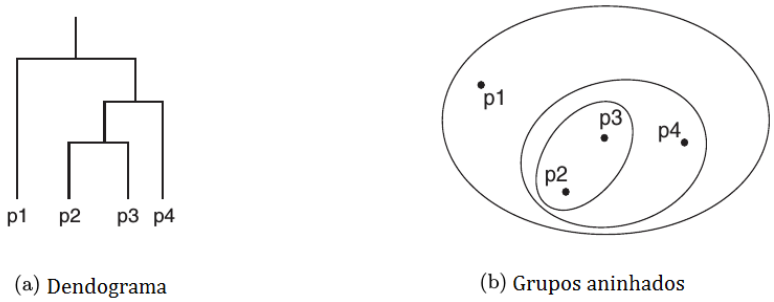


Figura 2.11. Agrupamento baseado em protótipo com elementos representativos circulados de preto.



(a) Dendograma (b) Grupos aninhados

Figura 2.12. Dendograma (a) e clusters aninhados (b) de um grupamento hierárquico (Fonte: Tan et al. [2006])

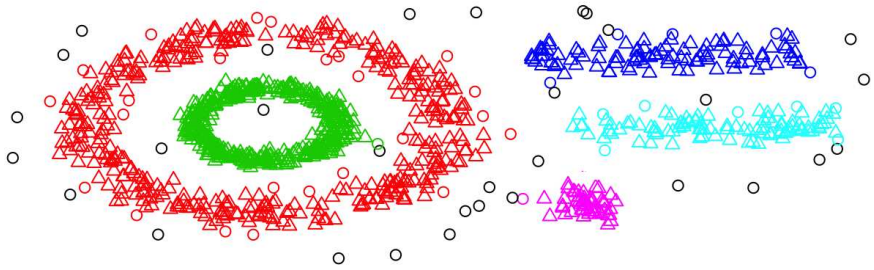


Figura 2.13. Exemplo de agrupamento usando técnicas baseadas em densidade.

Para este trabalho foi escolhido o algoritmo de agrupamento k-medoids, que é particional. Uma descrição mais aprofundada do k-medoids será feita na seção a seguir.

2.4.1 K-medoids

O k-medoids é um algoritmo de agrupamento e baseado em protótipo e particional. Dado um conjunto de n objetos, um método particional constrói k partições dos dados onde cada partição representa um grupo e $1 \leq k \leq n$. Geralmente, métodos particionais adotam grupos exclusivos, ou seja, cada objeto deve pertencer a somente um grupo [Han et al., 2011].

Para dados de atributos contínuos, o protótipo é frequentemente um centroide que representa a média de todos os pontos do grupo. O algoritmo mais citado que utiliza a técnica de centroides é o k-means. Contudo, a média calculada por esse tipo de técnica é muito sensível a ruídos. Pontos extremos podem levar os centroides para regiões não desejadas, formando piores grupos. Por esse motivo, para este trabalho, foi escolhido o k-medoids, pois é mais robusto que o k-means [Sammut & Webb, 2011], pelo fato de os protótipos serem medianas (pontos que pertencem aos dados).

Enquanto raramente um centroide representa um ponto dos dados, no k-medoids, por definição, o protótipo deve ser um ponto pertencente aos dados [Tan et al., 2006]. Na Imagem 2.14, é exibido o posicionamento de um protótipo para o k-means e o k-medoids em um mesmo grupo de objetos. É notável que devido ao ruído (ponto mais afastados dos demais), no k-means, o centróide é arrastado pra uma região fora de onde estão a maioria dos objetos. Usando um elemento real dos dados como ponto representativo é possível evitar esse desvio.

Figura 2.14. Posicionamento dos protótipos para os algoritmos k-means e k-medoids em um mesmo conjunto de dados.

Obter um particionamento ótimo global é computacionalmente custoso, pois requer uma enumeração exaustiva de todas partições possíveis. Aplicações como os algoritmos k-means e k-medoids adotam métodos heurísticos de caráter guloso para melhorar progressivamente a qualidade do agrupamento e alcançar um ótimo local [Han et al., 2011]. O PAM (Partitioning Around Medoids) [Kaufman & Rousseeuw, 1990], é uma técnica que implementa o k-medoids. A ideia básica é mostrada no Algoritmo 1. A entrada para o Algoritmo 1 é um número k de grupos e uma base de dados D contendo n objetos. A saída do programa é o conjunto de objetos agrupados em k grupos.

O primeiro passo do PAM é selecionar um ponto representativo inicial qualquer para cada grupo (Linha 1). Os demais pontos são atribuídos aos grupos levando em consideração a distância ao objeto representativo de cada grupo (Linha 3). Em sequência, é feita uma realocação iterativa que melhora o particionamento ao mover objetos de um grupo para outro (Linhas 4 a 8). O critério de parada é quando não houver mais trocas entre os elementos representativos (Linha 9). No fim do processo é esperado que objetos em um mesmo grupo sejam próximos, ou bem relacionados, e objetos de grupos diferente sejam dissimilares.

Algoritmo 1: *PAM, um algoritmo K-medoids baseado em objetos centrais.*

1: Escolhe aleatoriamente k objetos em D para serem os protótipos iniciais;
2: **repeat**
3: Atribui cada objeto remanescente ao grupo com protótipo mais próximo;
4: Seleciona aleatoriamente um objeto não representativo, o_{novo};
5: Calcula o custo S de trocar um objeto representativo o_j com o_{novo}, $1 \leq j \leq k$;
6: **if** $S \leq 0$ **then**
7: Swap o_j com o_{novo} para formar um novo conjunto de k objetos representativos;
8: **end if**;
9: **until** Não haver mais trocas

2.4.2 Coeficiente de Silhueta

Existem métricas utilizadas para mensurar a qualidade de um agrupamento. Esse tipo de medida é bastante útil para entender a tendência dos dados quando agrupados, determinar um número correto de grupos, ou avaliar diferentes agrupamentos e verificar qual é o melhor para um determinado contexto [Tan et al., 2006]. Quando não se tem um conhecimento exato sobre a natureza dos dados, para sugerir um número adequado de grupos, pode-se usar tais métricas executando o algoritmo diversas vezes e avaliar qual o número adequado de partições para um conjunto de dados específico.

O coeficiente de silhueta é um método chamado de intrínseco [Han et al., 2011]. Tais métodos avaliam um agrupamento ao examinar o quanto os grupos estão separados e como cada grupo está compacto. É uma medida que verifica coesão e separação dos grupos, sendo baseada na diferença entre as médias das distâncias entre pontos de grupos distintos e no mesmo grupo.

Seja uma base de dados D de n objetos e suponha que D está particionado em k grupos, $(C_1, ..., C_k)$. Para cada objeto $x_i \in D$, seu coeficiente de silhueta s_i é dado como:

$$s_i = \frac{b(x_i) - a(x_i)}{max\{a(x_i), b(x_i)\}}, \tag{2.1}$$

onde $a(x_i)$ (Equação 2.2) é a média das distâncias entre $x_i \in C_l (1 \leq l \leq k)$ e todos os objetos do grupo a que x_i pertence. Quanto menor este valor, mais compacto é o grupo.

$$a(x_i) = \frac{\sum_{x_j \in C_l, x_i \neq x_j} dist(x_i, x_j)}{|C_l| - 1} \tag{2.2}$$

Já $b(x_i)$ (Equação 2.3) é a menor média das distâncias de x_i em relação a pontos em outros grupos. Quanto maior o valor de $b(x_i)$ significa que x_i está mais afastado dos outros grupos.

$$b(x_i) = min_{C_m:1\leq m\leq k, m\neq l}\{\frac{\sum_{x_j\in C_m} dist(x_i, x_j)}{|C_m|}\} \tag{2.3}$$

O valor s_i de um ponto está no intervalo [-1,+1]. Um valor perto de 1 indica que x_i está mais próximo dos pontos em seu grupo do que em outros grupos. Um valor próximo a zero indica que o ponto está na fronteira entre grupos. Finalmente, um ponto perto de -1 indica que x_i está mais próximo de outros grupos do que de seu próprio grupo e, portanto, esse ponto precisa ser realocado.

Para medir a qualidade de um grupo em um agrupamento, pode-se calcular a média do coeficiente de silhueta de todos os objetos daquele grupo. Para medir a qualidade do agrupamento, seu coeficiente de silhueta médio é definido como a média s_i sobre todos os pontos da base de dados, onde um valor próximo a +1 indica um bom agrupamento:

$$SC = \frac{1}{n}\sum_{i=1}^{n} s_i \tag{2.4}$$

Observando a Figura 2.15, no quadro à esquerda está o particionamento de alguns objetos de dados em um agrupamento particional com cinco grupos. No quadro à direita da Figura 2.15 estão os valores de coeficiente de silhueta para os dados, coloridos de acordo com o grupo em que estão contidos. Tomando como exemplo o grupo 1 e o grupo 4, é possível notar que o coeficiente de silhueta consegue mensurar a forma como os dados de cada grupo estão distribuídos. No grupo 1, os objetos além de próximos, estão também afastados dos demais grupos. Isso garante um alto valor médio de coeficiente de silhueta para o grupo 1. Já o grupo 4, além de possuir elementos dispersos, pouco compactos, apresenta também objetos muito próximos do grupo 2, diminuindo e em alguns casos até negativando seu coeficiente de silhueta.

2.5 Trabalhos relacionados

Vários trabalhos têm sido propostos para o estudo de interações proteína-ligante. São diversas abordagens que trabalham com a caracterização de contatos e visualização de estruturas e interações sendo que algumas também utilizam modelagem em grafos.

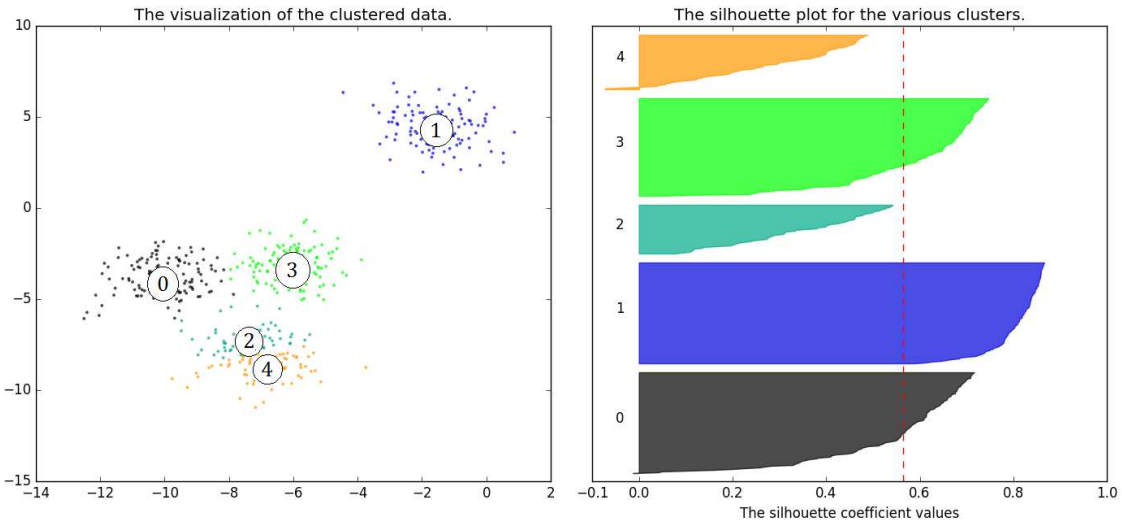


Figura 2.15. Análise do coeficiente de silhueta para um agrupamento particional em dados amostrais com 5 grupos.

Ferramentas como o LigPlot [Wallace et al., 1995] e LigPlot+ [Laskowski & Swindells, 2011] lidam com a geração de esquemas bidimensionais (2D) de interações proteína-ligante a partir de dados estruturais 3D. Essa esquematização deve-se à simplificação visual que a perspectiva 2D pode oferecer. O ProBiS-ligands [Konc & Janežič, 2014] utiliza modelagem baseada em grafos com aplicações de algoritmos de busca para identificar ligantes capazes de ligar a uma proteína consultada. O HydroPaCe [Gonçalves-Almeida et al., 2012] utiliza cálculo de contatos para definir interações em nível atômico na interface proteína-ligante. Os contatos que representam interações apolares na interface de ligação proteína-ligante são modelados como grafos, e algoritmos baseados em grafos são utilizados na tentativa de detectar caminhos hidrofóbicos conservados na estrutura.

Existem abordagens também a fim de somente detectar interações na interface prteína-ligante e caracterizá-las. O STING [Mancini et al., 2004], por exemplo, é uma ferramenta que analisa e calcula contatos em proteínas em termos de interações atômicas. O PLIP [Salentin et al., 2015] também caracteriza interações entre proteína e ligante, detectando contatos relevantes em estruturas 3D de complexos e disponibilizando sua visualização. O LigDig [Fuller et al., 2015] também trabalha com visualização, apresentando uma ferramenta que auxilia na busca de características referentes às proteínas, os ligantes e suas interações.

O GReMLIN contribuirá para o estado da arte com uma nova estratégia que combina modelagem e algoritmos baseados em grafos, para encontrar estruturas conservadas nas interfaces de interação proteína-ligante. Nossa ferramenta encontra

padrões estruturais, em nível atômico, que podem ser úteis para elucidar características relevantes nas interações proteína-ligante.

Capítulo 3

MÉTODOS

Para encontrar estruturas conservadas em interações proteína-ligante, é executada uma sequência de programas que envolvem desde a modelagem dos arquivos estruturais de proteínas, até utilização de algoritmos para encontrar os padrões conservados. A estratégia foi dividida em três etapas principais, sendo cada etapa também dividida em funções menores, com o objetivo de tornar o processo inteligível. Apresentamos na Figura 3.1 os três grande blocos do GReMLIN, assim como as tarefas executadas em cada uma delas.

A primeira etapa, denominada *Modelagem* (Figura 3.1(a)), é responsável por transformar os dados de entrada, que são os arquivos estruturais de proteínas, em

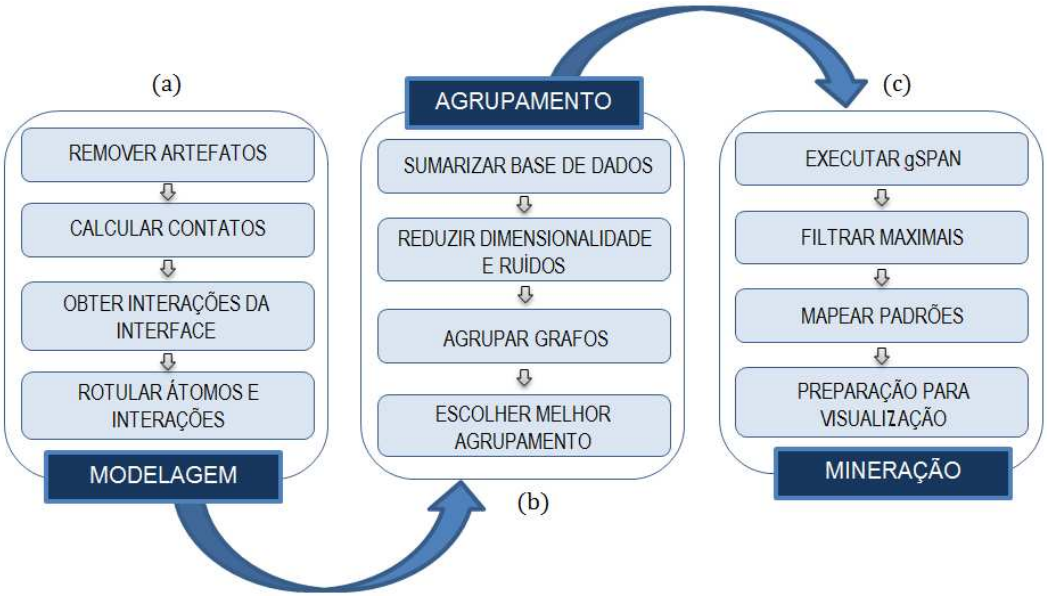


Figura 3.1. Fluxo de execução do GReMLIN.

grafos que representam as interfaces físicas de ligação entre os átomos da proteína e do ligante. A etapa subsequente, *Agrupamento* (Figura 3.1(b)), é responsável por fazer uma análise de agrupamento na base de grafos oriunda da etapa de Modelagem e retornar os grafos particionados em grupos disjuntos. A etapa de *Mineração* (Figura 3.1(c)), terceira e última, executa um algoritmo de mineração de subgrafos para extrair padrões em cada grupo gerado na etapa anterior.

Nas próximas seções, cada etapa será apresentada de forma mais aprofundada, assim como seus subprocessos, abordando as técnicas e modelagens utilizadas, sua importância, e como elas interagem no decorrer da estratégia.

3.1 Modelagem

A etapa de modelagem é responsável por transformar os complexos estruturais de proteínas e ligantes em grafos. Os arquivos são obtidos do Protein Data Bank [Berman et al., 2000]. Tais arquivos podem ser de uma família ou um grupo de proteínas que compartilham características estruturais semelhantes. Cada arquivo de entrada possui um identificador, denominado PDBid, e representa uma proteína e suas cadeias em complexo com ligantes.

Primeiramente é feito o pré-processamento para eliminar características indesejáveis na base de dados. Uma tarefa inicial é remover átomos de hidrogênio existentes nos arquivos, pois a presença do mesmo é inconsistente, variando de um arquivo para outro. Existem várias abordagens para resolução estrutural das biomoléculas, sendo que algumas conseguem identificar hidrogênios na estrutura e outras não [Garbuzynskiy et al., 2005]. Portanto, é viável a remoção total dos hidrogênios para que este não interfira no processo de mineração. Em seguida, é feita a remoção de artefatos de cristalografia. Ligantes com seis átomos ou menos são considerados artefatos e são removidos, assim como feito em [Pires et al., 2013].

Cada arquivo PDBid é particionado em termo das cadeias da sua respectiva proteína. Para cada cadeia é criado um novo arquivo, denominado PDBid-cadeia, que possui como identificadores seu id no PDB e sua respectiva cadeia. Além disso, arquivos sem a presença de ligantes são desconsiderados, já que não faz sentido procurar padrões em interações inexistentes. Nesse momento, a base de dados, constituída pelos arquivos resultantes, se encontra pronta para iniciar sua transformação em grafos.

3.1.1 Construindo os grafos

A modelagem inicia-se com a extração das coordenadas tridimensionais dos átomos da proteína em complexo com o ligante. Com a posição de cada átomo no espaço é possível realizar o cálculo de contatos através do diagrama de Voronoi e da triangulação de Delaunay [Klein & Lee, 2014], que é uma abordagem de corte no espaço Euclidiano e que evita oclusão. Dado um conjunto de objetos p, que exercem influência sobre seu espaço de vizinhança S, seleciona-se os pontos $z \in S$ em que a influencia de p é mais forte e cria-se uma aresta entre z e p (Figura 3.2). Na prática, os pontos são os átomos e a influência corresponde às interações intra ou intermoleculares entre eles. A proximidade dos átomos, usando a distância Euclidiana como métrica, define a existência ou não de uma interação.

Figura 3.2. Exemplo de uma Triangulação de Delaunay de 10 pontos no plano Euclidiano. Fonte: Klein & Lee [2014].

Para construir os grafos serão utilizados somente os átomos da interface de ligação entre a proteína e o ligante. Denomina-se interface o conjunto de átomos presentes na camada mais externa, tanto da proteína quando do ligante, e que interagem entre si através de interações intermoleculares. Na Figura 3.3 está ilustrado um esquema de como um grafo é construído.

A partir do conjunto de arestas calculadas pela triangulação de Delaunay, são filtradas somente arestas que conectam um átomo pertencente à proteína com um átomo que pertence ao ligante. As arestas resultantes constituem os grafos, sendo que cada grafo oriundo desse processo é um grafo bipartido. Um grafo bipartido é formalmente definido como um grafo $G(P, L, E)$, cujos vértices estão divididos em dois conjuntos disjuntos, P e L, de forma que toda aresta em E conecta um vértice em P a um vértice em L. Assim, o conjunto de vértices P representam átomos da proteína, o conjunto de vértices L representa os ligantes e o conjunto de arestas E representam as interações entre o átomo da proteína e o átomo do ligante.

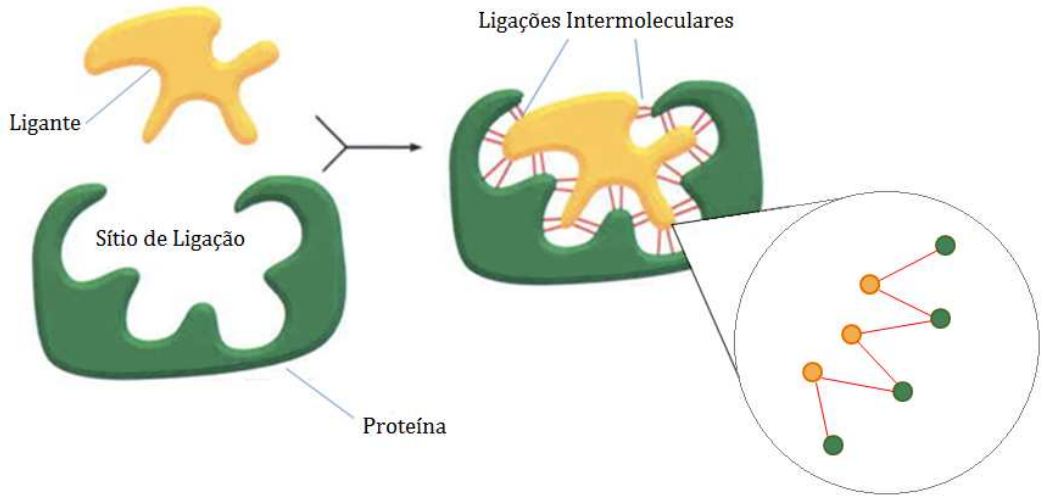


Figura 3.3. Esquema de construção dos grafos a partir dos átomos da interface de contato da proteína e do ligante.

Na Figura 3.4 está representado um dos grafos modelados para a proteína 1IL5:A (PDBid 1IL5 e cadeia A) ligada ao ligante DDP através de interações hidrofóbicas. Observando a organização dos vértices, é possível notar que o grafo é bipartido. Os átomos do lado da proteína estão representados por vértices azuis, enquanto os átomos do ligante estão representados por vértices verdes. As interações hidrofóbicas entre os pares de átomos são representadas pelos segmentos que conectam sempre vértices de conjuntos distintos (um da proteína e um do ligante).

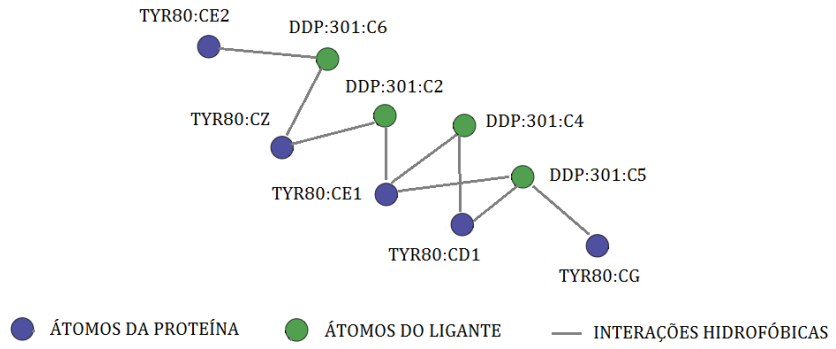


Figura 3.4. Esquema da modelagem de um grafo para a proteína 1IL5:A e seu ligante DDP.

Diferentes regiões da proteína podem estabelecer interações com um ou mais ligantes. Para atender essa característica, foi feito o cálculo de componentes conexas para cada arquivo de entrada. Assim, cada PDBid-cadeia corresponderá a um ou

mais grafos conectados. Como resultado, cada componente conexa será um grafo de entrada para a etapa de mineração e cada grafo pode representar interações proteína-ligante para um ou mais ligantes.

Na Figura 3.5 está representado um exemplo de componente conexa. O Ligante 1 e o Ligante 2 estão interagindo em uma região próxima da proteína. Portanto, compartilham um mesmo átomo da interface em seus grafos. Caso estas estruturas sejam consideradas separadamente no processo de mineração, informações serão perdidas, como por exemplo, o subgrafo $\{a, b, c\}$ seria um possível padrão desconsiderado. Portanto, usar componentes conexas como grafos para a mineração é mais apropriado que considerar interações de cada ligante como um grafo para uma entrada específica do PDB (PDBid). Finalmente, a estrutura dos grafos da base de dados está definida e o próximo passo é rotular seus vértices e arestas.

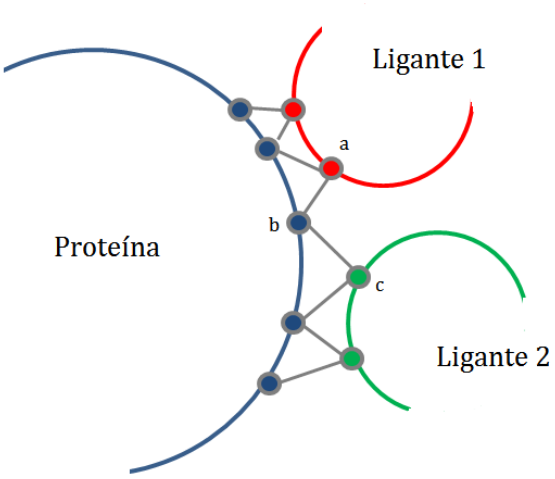


Figura 3.5. Esquema de uma componente conexa formada pela combinação de grafos de ligantes distintos.

3.1.2 Rotulando vértices e arestas

Além da estrutura, utilizamos atributos relacionados com as propriedades de interação dos átomos para caracterizar os grafos. Vértices e arestas são rotulados de acordo com suas propriedades físico-químicas. Os vértices da proteína foram rotulados com as propriedades *positivo*, *negativo*, *hidrofóbico*, *doador* e *aceptor* de acordo com [Gonçalves-Almeida et al., 2012; Sobolev et al., 1999] . Os vértices dos ligantes foram rotulados com os mesmos tipos, porém usando o software Pmapper da Chemaxon (Pmapper 5.3.8, 2010, Chemaxon 2) em pH 7.0. As arestas foram rotuladas como: *aromático*, *ligação de hidrogênio*, *hidrofóbico*, *repulsivo* e *ponte salina*, levando

em consideração o tipo dos vértices e um critério de distância em Å(ångström) pré-definido de acordo com [Mancini et al., 2004], como pode ser observado na Tabela 3.1.

Tabela 3.1. Critérios para rotular uma interação entre átomos da proteína e do ligante.

Tipo de Interação	Tipo do Átomo	Distância Mínima	Distância Máxima
Aromático	2 átomos aromáticos	1.5 Å	3.5 Å
Ligação de Hidrogênio	1 aceptor e 1 doador	2.0 Å	3.0 Å
Hidrofóbico	2 átomos hidrofóbicos	2.0 Å	3.8 Å
Repulsivo	2 átomos com a mesma carga	2.0 Å	6.0 Å
Ponte Salina	2 átomos com cargas opostas	2.0 Å	6.0 Å

Vértices e arestas podem receber múltiplos rótulos, pois estes frequentemente carregam múltiplas propriedades. Tomando como exemplo novamente o grafo oriundo da interação entre a proteína 1IL5:A com o ligante DDP na Figura 3.6, porém agora com os rótulos, é possível observar que os vértices do lado da proteína possuem rótulos múltiplos (aromático/hidrofóbico). Com os rótulos fixados, temos a base de dados constituída por grafos, bipartidos e rotulados, que modelam a interface de interações entre proteínas e ligantes.

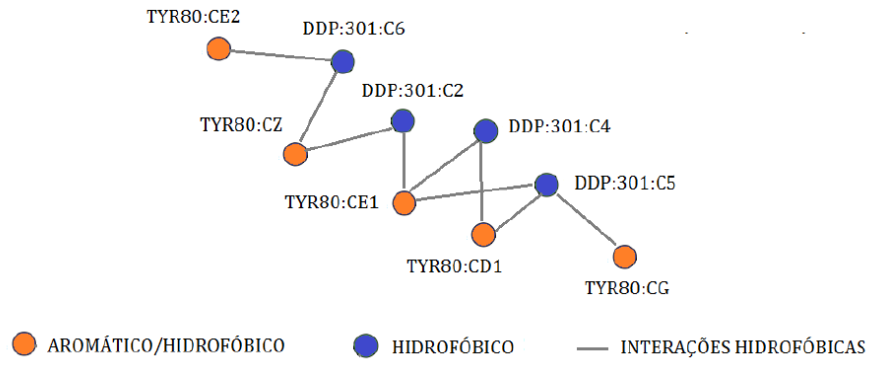


Figura 3.6. Esquema da modelagem de um grafo rotulado com propriedades físico-químicas.

3.2 Agrupamento

Antes de iniciar a mineração de subgrafo frequente na base de dados, foi feita uma análise de agrupamentos para organizar os grafos em grupos disjuntos. Como o objetivo é encontrar estruturas conservadas nas interações, é possível que estas estruturas se conservem em grafos com características semelhantes. Portanto, trabalhar

na busca de padrões em grupos de elementos com alta similaridade aparenta ser o caminho para encontrar fragmentos estruturais conservados mais coerentes.

Para agrupar os grafos criou-se uma matriz de atributos com a finalidade de caracterizá-los. A base de grafos é sumarizada na matriz construída, denominada *matriz de contagem*, onde cada linha representa um grafo da base e cada coluna define os tipos de interações presentes neste grafo. Cada uma das entradas da matriz enumera a quantidade de interações presentes em um determinado objeto.

Na Tabela 3.2 apresentamos um exemplo fictício de uma matriz de contagem. Na primeira coluna da matriz estão os grafos da base de dados, nas demais colunas encontra-se, em cada coluna, um valor que denota a quantidade de interações presentes no grafo representadas pelos rótulos das extremidades de cada aresta.

Tabela 3.2. Exemplo de uma matriz de contagem

Grafo Rotulado	Acceptor:Acceptor/Doador	Acceptor:Doador	Acceptor:Aromático/Doador
GRAFO 1	2	0	0
GRAFO 2	2	0	1
GRAFO 3	2	1	0

Com auxílio da Imagem 3.7 para ilustrar a base de grafos da tabela acima, é possível observar como as interações são dispostas nas entradas da matriz de contagem. Por exemplo, o $GRAFO2$ possui 2 interações do tipo *Acceptor* $\longleftrightarrow$ *Acceptor/Doador* e uma interação do tipo *Acceptor* $\longleftrightarrow$ *Aromático/Doador*.

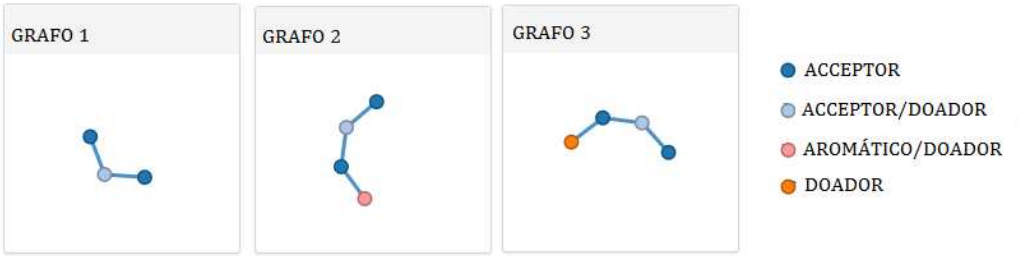


Figura 3.7. Base de grafos fictícia para exemplificar a matriz de contagem.

Foi executado sobre a matriz de contagem construída um algoritmo de Decomposição de Valores Singulares (SVD) [Klema & Laub, 1980], com o objetivo de compactar a matriz, reduzindo seu número de dimensões e removendo ruídos presentes nos dados.

Para agrupar os grafos foi usado o algoritmo k-medoids [Claude Sammut, 2010], apresentado na Seção 2.4.1. O objetivo é dividir a base de dados em grupos disjuntos, deixando elementos com alta similaridade em um mesmo grupo e elementos

dissimilares em grupos distintos. Ao receber como entrada a matriz de contagem da base de grafos e um número pré-definido de grupos, o k-medoids atribui cada grafo a um grupo baseando-se nas características que cada um possui. Assim, grafos que compartilham interações semelhantes tendem a ficar em um mesmo grupo, proporcionando maior consistência no momento em que os padrões forem extraídos em cada grupo individualmente.

Como o k-medois é sensível às configurações iniciais, neste caso o número de grupos (k), o algoritmo é executado para todos os possíveis valores de k. Exemplificando, se existe uma base com n grafos, desconsiderando os agrupamentos triviais com $k = 1$ e $k = n$, serão executados agrupamentos com k variando de 2 a $n - 1$. Para cada agrupamento obtido é calculado sua respectiva média do *coeficiente de silhueta* (Seção 2.4.2), uma métrica que quantifica a qualidade de um agrupamento em termos das distâncias entre elementos de um mesmo grupo e em grupos distintos [Pang-Ning et al., 2006]. Dentre todos os agrupamentos, é selecionado aquele que possui maior coeficiente, isto é, o agrupamento mais coeso.

Realizando o agrupamento dos dados, o próximo passo é minerar os padrões. Além de facilitar a etapa de mineração, entregando para a mesma conjuntos de grafos que compartilham semelhanças, ter a base de dados dividida em grupos auxilia na análise dos resultados, possibilitando inferir características que ajudam no reconhecimento das interações entre proteína e ligante de um mesmo grupo.

3.3 Mineração

Minerar os padrões dos grupos de grafos é a ultima tarefa da estratégia. Aqui, cada grupo originado na etapa de agrupamento é submetido a um algoritmo de mineração de subgrafo frequente, que extrai de um conjunto de grafos subestruturas que aparecem frequentemente (mais detalhes na Seção 2.3.1).

O algoritmo utilizado para extrair subgrafos frequentes da base de grafos foi o gSpan [Yan & Han, 2002b], algoritmo mais citado na literatura [Jiang et al., 2013]. Para executar o gSpan, além de uma base de dados, é necessário um valor de *suporte* (μ). Fornecido pelo usuário, μ é um valor decimal entre 0 e 1 que representa um valor mínimo de suporte para os padrões, ou seja, um percentual mínimo referente à base de grafos em que o padrão deva ocorrer.

Um valor ideal para μ deve ser definido empiricamente. Para isso, em cada grupo de grafos, é executado o algoritmo de mineração de subgrafo frequente com μ variando entre 0.1 e 1.0, com intervalos de 0.1. Assim podemos analisar os padrões

olhando de vários pontos de vista diferentes ou, neste caso, suportes diferentes. Os subgrafos encontrados pelo gSpan representam estruturas conservadas nas interfaces entre proteína e ligante.

Devido à grande quantidade de padrões que podem ser produzidos em uma execução do gSpan, foi feito um filtro nos subgrafos frequentes, conservando somente os padrões maximais e eliminando os demais. Em um grupo, um subgrafo é maximal quando este não está contido em nenhum outro subgrafo, como foi definido na Seção 2.3.1. Em outras palavras, padrões menores que estão contidos em outros maiores são eliminados. A intenção de extrair subgrafos maximais é diminuir o número de dados a serem analisados sem que a informação seja perdida, já que a partir da fragmentação dos subgrafos maximais é possível chegar aos padrões menores.

Para que as estruturas frequentes mineradas pela estratégia sejam preparadas para visualização é necessário um mapeamento dos padrões sobre os grafos de entrada originais. O gSpan encontra padrões frequentes na base de dados, porém não especifica onde exatamente os padrões se encontram nos grafos de entrada. Para que as estruturas frequentes sejam indicadas corretamente sua localização nos grafos originais é necessário fazer uma operação de isomorfismo de subgrafo entre os padrões resultantes e os grafos de entrada. A Figura 3.8 mostra um padrão sendo mapeado para um grafo em que o mesmo está contido.

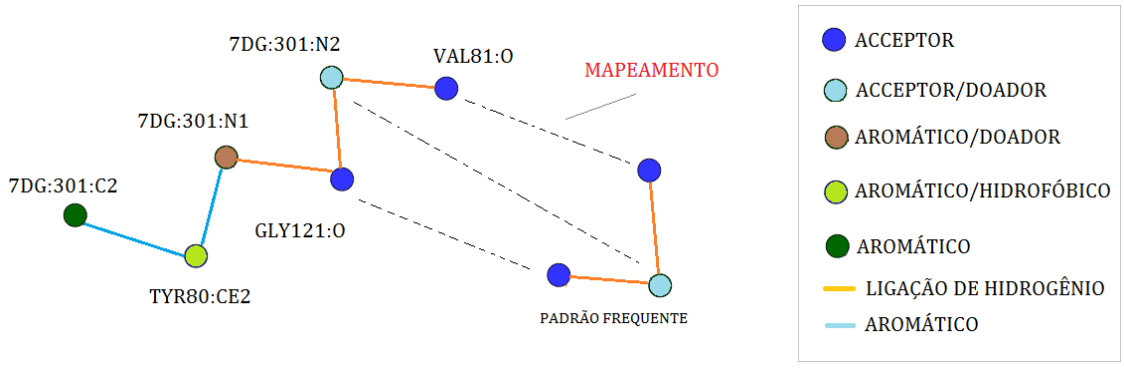


Figura 3.8. Exemplo de mapeamento de um padrão frequente para um grafo de entrada da base de grafos.

Esta operação é executada para todas subestruturas relevantes e todos os grafos de entrada. Para fazer o mapeamento foi utilizado o algoritmo de isomorfismo de subgrafo VF2 [Cordella et al., 2004] implementado em linguagem C na biblioteca *igraph 0.7.1*[1].

[1]http://igraph.org/redirect.html

Capítulo 4

EXPERIMENTOS e RESULTADOS

Para validar a estratégia proposta nesse trabalho e mostrar sua aplicação em um cenário real, foram coletados dados estruturais de proteínas em complexo com ligantes do PDB. Com estes dados foram criadas duas bases de dados, uma referente à CDK2 humana (cyclin dependent kinases), e a outra referente à ricina, ambas de relevância biológica, porém com diferentes características.

4.1 Dados da CDK2

CDKs são uma família de proteínas quinases envolvidas na progressão e transcrição do ciclo celular. Sua desregulação tem sido associada ao câncer e doenças neurodegenerativas, tornando-as alvos importantes na descoberta de fármacos [Johnson, 2009; Schonbrunn et al., 2013].

Estudos cristalográficos permitiram um entendimento detalhado do mecanismo molecular da CDK2. Seu núcleo catalítico é composto de múltiplos subdomínios conservados encontrados em todas as proteínas quinases [Silva et al., 2009]. Sua estrutura é composta por uma hélice C, um grande domínio C-terminal, constituída predominantemente por hélices-alfa e um domínio N-terminal formado por folhas-beta (Figura 4.1). O sítio de ligação situa-se na interface domínio-domínio.

A base de dados da CDK2 é composta por 73 entradas do PDB com sequências idênticas acopladas a diferentes ligantes. Essa base de dados base de dados foi derivada de [Schonbrunn et al., 2013], onde resíduos relevantes para interação proteína-ligante foram experimentalmente determinados. Os padrões obtidos utilizando o GReMLIN foram comparados com resíduos e átomos relevantes para a

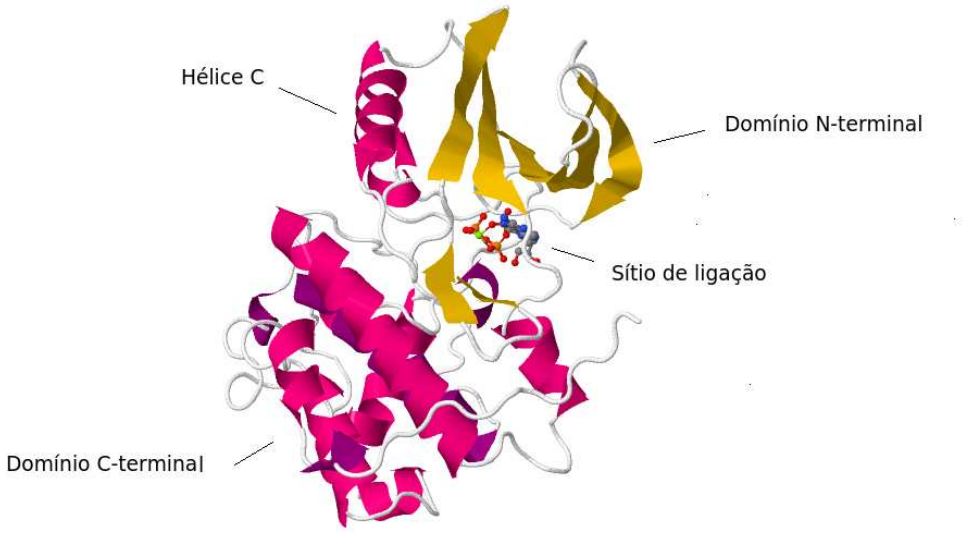


Figura 4.1. Estrutura da CDK2 (PDB: 1HCK).

interação das CDKs com ligantes experimentalmente determinados em [Schonbrunn et al., 2013].

4.2 Dados da Ricina

O conjunto de dados da ricina é composto por 29 entradas do PDB, que compartilham sequência com semelhança maior ou igual a 50% com o template 2AAI.A da ricina no PDB. A ricina é uma toxina natural encontrada na semente da mamona para proteger a planta de pestes. Sua estrutura proteica e toxidade tem sido estudada extensivamente nas ultimas décadas [HOFFMANN et al., 2007].

A estrutura da proteína ricina é globular e composta por duas subunidades, a cadeia A citotóxica e a cadeia B receptora de ligações (Figura 4.2) . As cadeias A e B são unidas por uma ligação covalente de dissulfeto. A cadeia A é uma toxina inativadora de ribossomos, que inibe a síntese de proteínas e causa a morte da célula. Já a cadeia B, funciona como transportadora, capaz de ligar-se a carboidratos da superfície da membrana celular causando a entrada da cadeia A no citosol. Separadamente, a cadeia A não é tóxica, pois ela é incapaz de entrar na célula sozinha [HOFFMANN et al., 2007].

A ricina tem um apelo político e militar, considerada como um potencial agente de armas químicas de interesse para terroristas [Chen et al., 2015]. Além disso, há um grande interesse comercial para o uso dos resíduos (torta) após a extração do óleo

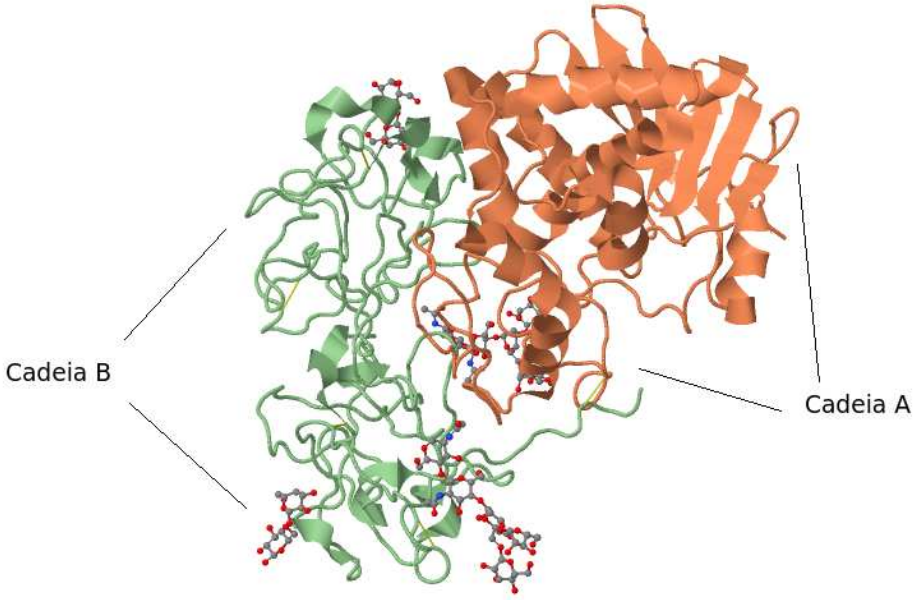


Figura 4.2. Estrutura da ricina (PDB: 2AAI).

da mamona. A torta possui alto valor proteico, o que a tornaria viável para ração animal. Porém, o uso da torta como alimento não é viável devido a presença dos elementos tóxicos e a falta de tecnologia praticável para seu processamento [Campos et al., 2016].

Entre os dados de teste obtidos do PDB, a ricina é a mais complexa. Os dados estruturais da ricina são escassos, o que é desafiador para algoritmos de mineração de dados. Além disso, nestes dados temos diferentes proteínas interagindo com diversos ligantes, o que torna o cenário mais realista, pois quando um especialista de domínio tem uma base de dados de interesse, ela é frequentemente composta por proteínas similares, ou seja, não são idênticas. Com os padrões resultantes do GReMLIN foi possível comparar átomos e resíduos relevantes para a interação das ricinas com ligantes experimentalmente determinados por [Ho et al., 2009].

4.3 Análise de agrupamentos

Durante a etapa de modelagem, os dados estruturais dos arquivos referentes à CDK2 e ricina foram transformados em grafos. Para os dados da ricina foram construídos 181 grafos conexos, numerados de 0 a 180, que representam as interfaces de interação proteína-ligante, enquanto para os dados da CDK2 foram criados 339 grafos conexos numerados de 0 a 338.

Antes de executar o algoritmo de mineração de subgrafo frequente sobre os grafos, estes são agrupados de acordo com suas similaridades e, para isso, é necessário a construção da matriz de contagem descrita na Seção 3.2. Em cada conjunto de dados é enumerado, para cada grafo, os tipos de interações ocorrentes. Assim, os grafos são os objetos dos dados (linhas), e os tipos de interações são os atributos dos objetos (colunas). Para a CDK2 gerou-se uma matriz com 339 linhas e 15 colunas, enquanto para ricina gerou-se uma matriz com 181 linhas e 22 colunas.

O próximo passo é utilizar o método de decomposição de valores singulares (SVD) para melhorar a qualidade dos dados com redução de dimensionalidade e remoção de ruídos. Seja d o parâmetro que define o número de dimensões da matriz resultante, o SVD foi executado com d variando de 1 a n, onde n representa o número de colunas da matriz inicial. Para cada uma das matrizes resultantes é executado o algoritmo de agrupamento k-medoids.

Para os dados da CDK o melhor coeficiente de silhueta médio obtido foi 1.0. Todas as matrizes resultantes do SVD foram agrupadas com o algoritmo k-medois, com d (número de dimensões da matriz de contagem) variando de 1 a 15, e os valores de k (número de grupos) variando de 2 a 338. Como resultado, o melhor coeficiente de silhueta médio foi para o agrupamento com $k = 5$ da matriz com $d = 1$ (Figura 4.3). Os grupos foram numerados de 1 a 5, com 158, 101, 75, 3 e 2 grafos respectivamente.

Os experimentos com os dados da ricina resultaram em um coeficiente de silhueta médio de 0.93. Foram testadas todas as matrizes que resultaram do SVD, com d variando de 1 a 22 e, para cada matriz de entrada, foi executado o agrupamento k-medoids variando as partições, isto é, os valores de k, entre 2 e 180. O melhor agrupamento obtido teve os parâmetros $d = 1$ e $k = 9$. Os valores de coeficiente de silhueta médios, de cada grupo e geral, podem ser observados na Figura 4.3. Os grupos da ricina foram numerados de 1 a 9 contendo 131, 16, 6, 6, 7, 3, 8, 2, 2 grafos respectivamente.

4.4 Análise dos padrões frequentes

Após dividir os grafos em grupos, para cada grupo, foi realizado uma mineração de subgrafo frequente para extrair os padrões conservados nos dados. O algoritmo gSpan foi executado sobre os dados com o suporte variando de 0.1 a 1.0 em intervalos de 0.1. Portanto, para cada grupo, existem padrões extraídos com 10 valores de suportes diferentes.

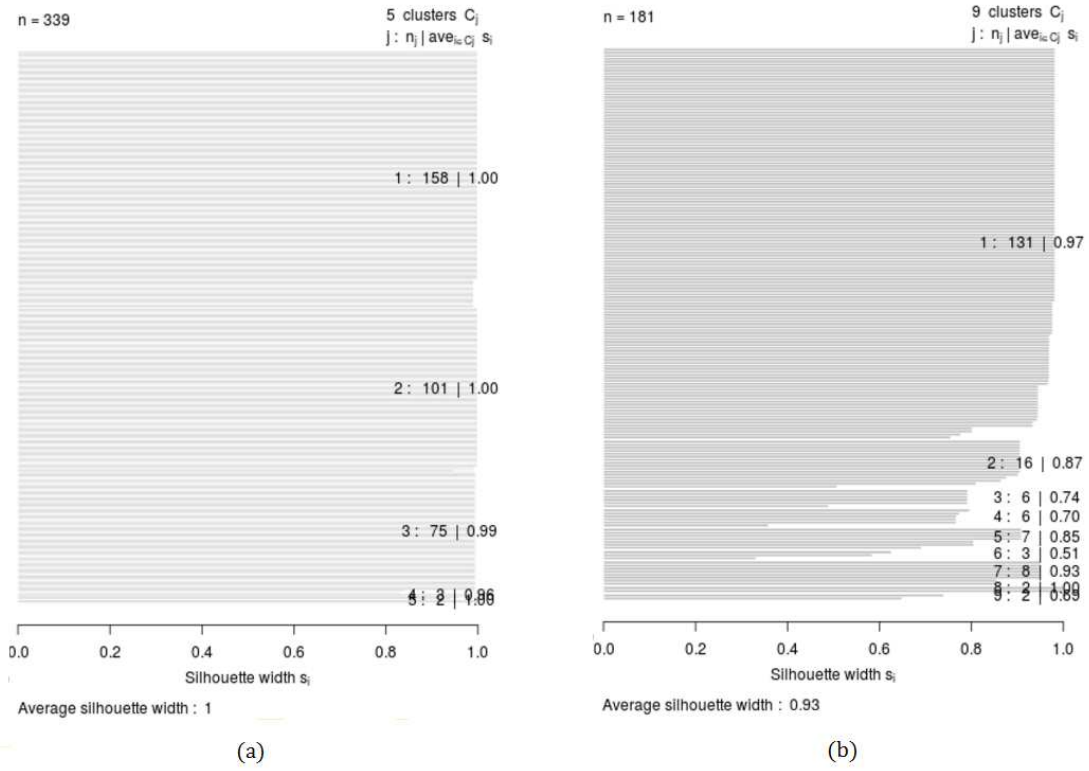


Figura 4.3. Coeficiente de silhueta médio para o agrupamento k-medoids com $k = 5$ para os dados da CDK2 (a) e para o agrupamento k-medoids com $k = 9$ para os dados da ricina (b).

A escolha do suporte ideal é uma tarefa empírica. Neste caso, foi criada uma tabela para verificação dos padrões de acordo com o valor do suporte. É desejável que estruturas frequentes apareçam nas interações proteína-ligante. Contudo, à medida que aumentamos o valor do suporte, alguns padrões são perdidos. Assim, existe um compromisso entre o suporte e o número de padrões.

Nas próximas seções serão descritas algumas subestruturas frequentes encontradas nos dados com o auxílio de figuras retiradas de uma ferramenta web de visualização denominada visGReMLIN[1], desenvolvida para auxiliar na análise dos padrões encontrados.

4.4.1 Análise dos padrões da CDK2

Observando os maiores valores de suporte na Tabela 4.1, para os dados da CDK, o valor mais apropriado para ser analisado foi 0.6. Ao observar a tabela é possível notar que os suportes 0.6 e 0.7 são valores altos com maior número de padrões,

[1] `http://homepages.dcc.ufmg.br/~sabrinas/visgremlin/index.html`

porém com o suporte 0.7 perdemos um padrão de tamanho 2 no grupo 1 destacado na Tabela 4.1 com borda vermelha. Portanto, foi escolhido o suporte 0.6 para análise.

Tabela 4.1. Número de padrões nos grupos da CDK2 (Suporte de 0.5 a 1.0).

Pattern size	Support: 0.5					0.6					0.7					0.8					0.9					1.0				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
2	1	1	1			1	1	1				1	1			1	1				1	1				1		1		
3				1					1					1					1					1						
4					2					2					2					2					2					2
5					1					1					1					1					1					1

O maior padrão encontrado para os grafos da base de dados da CDK2 foi um subgrafo de cinco vértices encontrado no grupo 5. As estruturas em que o padrão está contido foram o grafo 8 (3QQH:A, ligante X0A) e o grafo 10 (3QQJ:A, ligante X11). Como pode ser observado na Figura 4.4, para ambos os grafos, os vértices referentes ao lado da proteína foram HIS84:O (aceptor), LEU83:O (aceptor), e LEU83:N (doador). No grafo 8, os vértices encontrados do ligante foram XOA:303:N04 (aceptor/doador) e XOA:303:O21 (aceptor/doador), enquanto no grafo 10, os vértices do ligante foram X11:300:N8 (aceptor/doador) e X11:300:O15 (aceptor/doador).

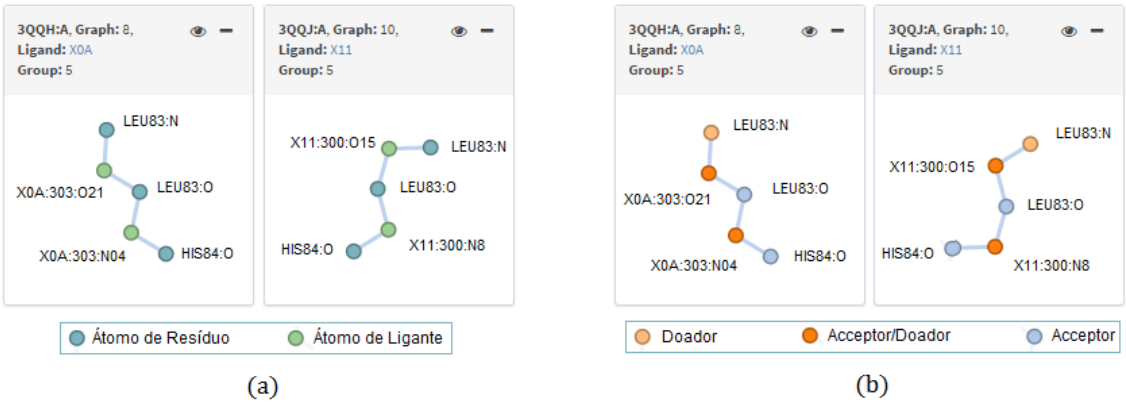


Figura 4.4. Padrões do grupo 5 com átomos diferenciados por moléculas (a) e por tipo (b).

Na Figura 4.5, está representado um esquema 3D dos padrões de interação para as proteínas 3QQH:A e 3QQJ:A. Os átomos da interface de interação proteína-ligante são exibidos, assim como segmentos que representam as interações entre os átomos.

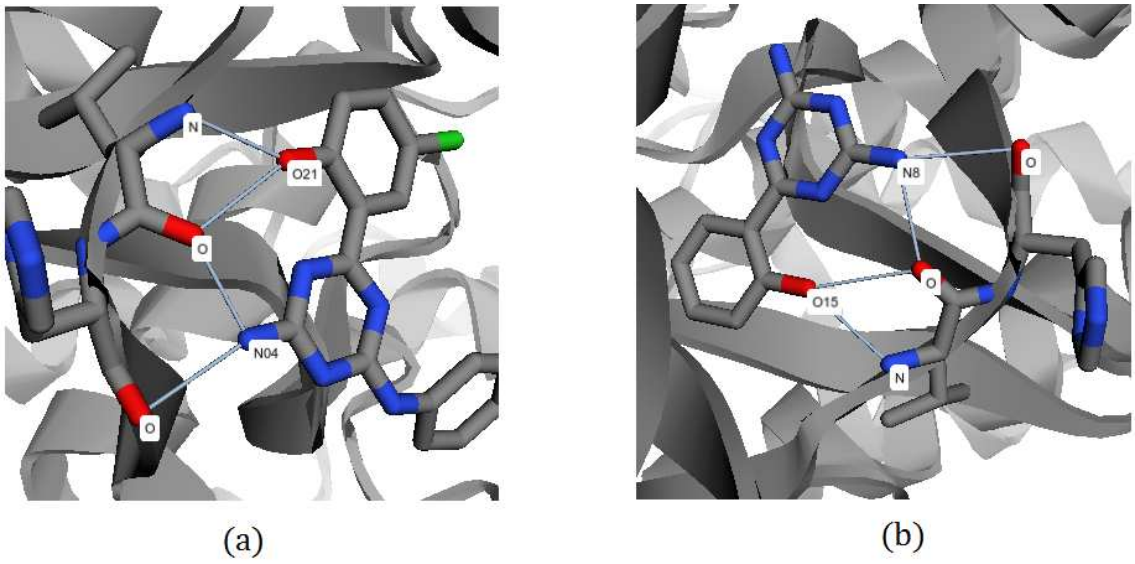


Figura 4.5. Representação em 3D dos padrões de interação do grupo 5 para CDK2.

Na Figura 4.6, são apresentadas as fórmulas estruturais dos ligantes X0A e X11 que interagem com a CDK (seus átomos de interação estão destacados em vermelho). É possível notar a existência de uma semelhança estrutural entre os ligantes, o que é um indício de que subestruturas frequentes de interação proteína-ligante também podem carregar implicitamente padrões estruturais das biomoléculas.

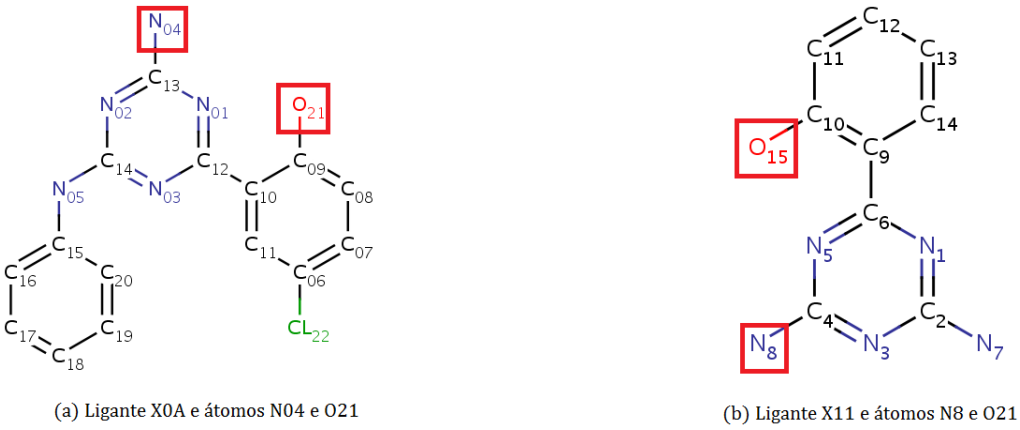


Figura 4.6. Estruturas dos ligantes X0A (a) e X11 (b)

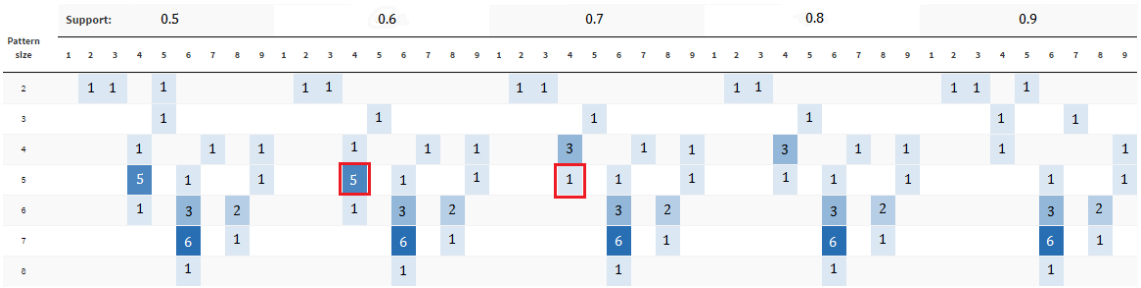
Nos grupos 1, 2 e 3 encontramos somente subgrafos com dois vértices e uma interação. O grupo 3 é composto por 75 grafos, sendo principalmente constituídos por pares de vértices aceptor e doador ligados por arestas rotuladas como ligação de hidrogênio. O grupo 2, composto por 101 grafos, é principalmente formado por

pares de átomos aceptor e aceptor/doador, consequentemente ligados por interações de hidrogênio, possuindo na maior parte das estruturas, átomos de ligantes rotulados como aceptor/doador e os átomos da proteína rotulados como aceptor. Além disso, os átomos de proteína são em sua maioria LEU83 e GLU81. O grupo 1 exibe principalmente átomos e interações hidrofóbicas, mas também possui alguns vértices hidrofóbicos/aromáticos, doador, doador/positivo, aceptor/doador e aceptor/aromático.

4.4.2 Análise dos padrões da ricina

O suporte escolhido para os dados da ricina foi 0.6 pois, observando a Tabela 4.2, pode-se notar que do suporte 0.6 ao 0.7 perdem-se 4 subgrafos de 5 vértices (destacados em vermelho) para o grupo 4. Considerando que o objetivo é obter o maior número de subestruturas com maior tamanho possível em cada grupo, o suporte 0.6 é uma escolha pertinente.

Tabela 4.2. Número de padrões nos grupos da Ricina (Suporte de 0.5 a 0.9).



O maior padrão encontrado nos dados da ricina é um subgrafo com 8 vértices, localizado no grupo 6 (3 grafos), nas estruturas de número 26 (1IL5:A, ligante DDP) e 101 (3RTI:B, ligante GAL). No grafo 26, todos os átomos de proteína são de TYR80 e no grafo 101 todos os átomos são de TRP37. Os átomos da proteína de ambos os grafos são rotulados como aromático/hidrofóbico e os átomos do ligante como hidrofóbico. Além disso, todas as arestas dos grafos representam interações hidrofóbicas. A Figura 4.7 mostra o padrão com 8 vértices (quadro com bordas verdes) e seu mapeamento nos grafos do grupo 6. No quadro (a) os vértices estão coloridos de acordo com suas propriedades físico-químicas, enquanto no quadro (b) os vértices estão representados de acordo com o tipo de molécula a que pertencem (proteína ou ligante).

Na Figura 4.8 temos uma representação 3D do padrão de interação entre a proteína 1IL5:A e o ligante DDP (Figura 4.9 (a)) e o padrão de interação entre a proteína 3RTI:B e o ligante GAL (Figura 4.9 (b)). Na Figura 4.9 foram destacados

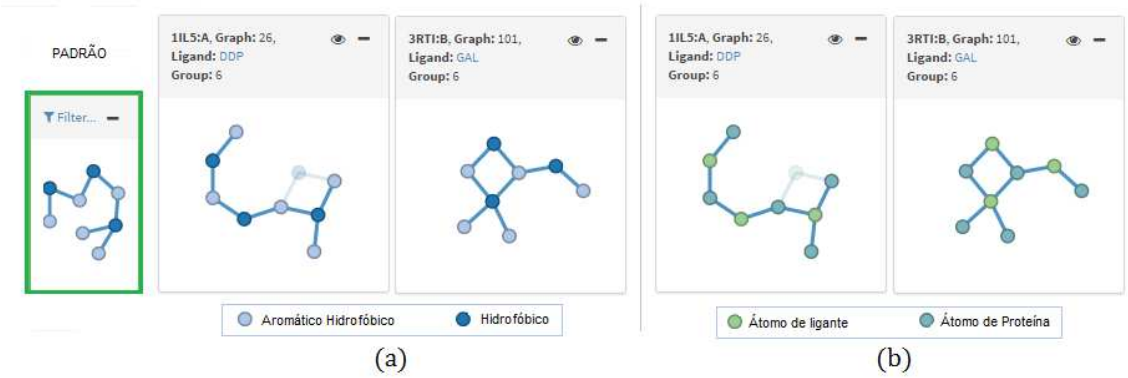


Figura 4.7. Padrão do grupo 6 com os vértices coloridos por propriedade (a) e tipo de molécula (b).

os átomos envolvidos no padrão em uma representação da fórmula estrutural dos ligantes.

No grupo 6 também existem 6 subgrafos com 7 vértices, 3 subgrafos com 6 vértices, e 1 subgrafo com 5 vértices, todos envolvendo átomos hidrofóbicos em relação ao ligante e átomos aromático/hidrofóbico para o lado da proteína, resultando em interações hidrofóbicas. No grupo 2, contendo 16 grafos, existe 1 subgrafo com dois vértices. Os átomos da proteína são TYR80, exceto para 1BR5:A que é de TYR123, rotulados com aromático/hidrofóbico, enquanto os vértices do ligante são rotulados como aromático. Consequentemente, interações aromáticas são frequentes

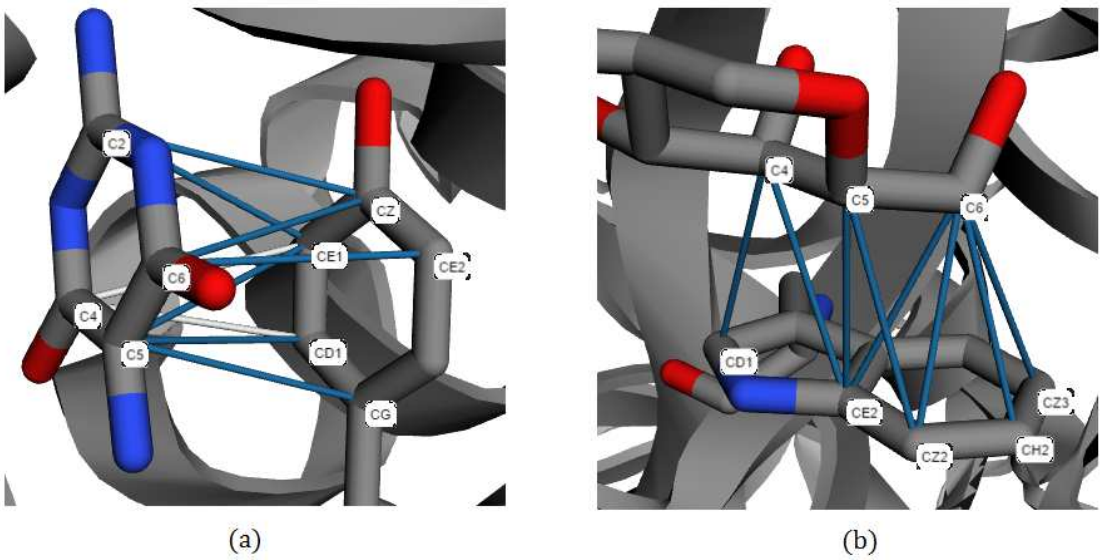
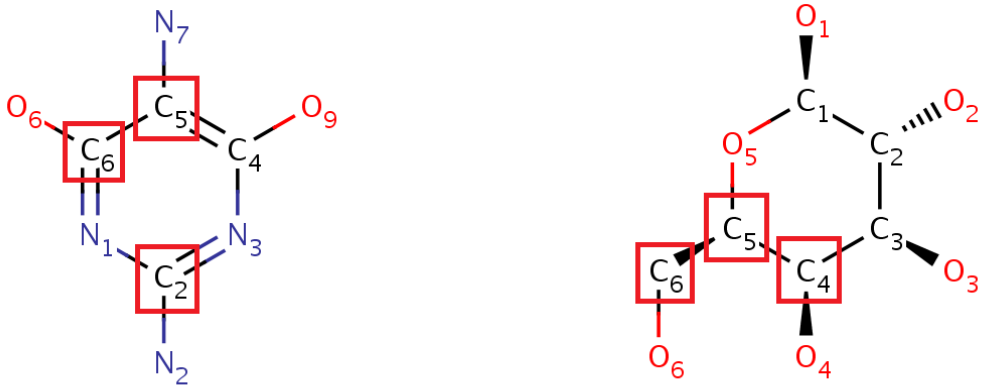


Figura 4.8. Esquema em 3D dos padrões de interação para o grupo 6 da Ricina.

no grupo.



(a) Ligante DDP, átomos C2, C5 e C6

(b) Ligante GAL, átomos C4, C5 e C6

Figura 4.9. Estruturas dos ligantes DDP e GAL

O maior conjunto de grafos dos dados da Ricina é o grupo 1 com 130 estruturas. É um grupo bastante heterogêneo, apresentando padrões somente para o suporte igual a 0.1. São 4 subgrafos frequentes com dois vértices e um subgrafo frequente com 3 vértices. Na Figura 4.10 apresentamos os grafos do grupo 1 em que o padrão de 3 vértices está contido. Neste padrão existem dois vértices referentes à proteína (GLY121 e VAL81) rotulados como aceptor e um vértice de ligante rotulado como aceptor/doador. Todas as interações são ligações de hidrogênio.

Fazendo uma análise visual das estruturas do grupo 1 é possível notar uma predominância de arestas rotuladas como ligação de hidrogênio. Contudo, seu número e forma não foram suficiente para torná-los frequente no grupo. Já era esperado que isso acontecesse, considerando que os dados da Ricina são heterogêneos, ou seja, compostos por entradas do PDB que possuem sequencia e estruturas bastante diversificadas entre si, ao contrário do conjunto de dados das CDKs, que possui entradas do PDB com sequências idênticas variando apenas o ligante.

4.5 Comparando padrões obtidos com GReMLIN com dados obtidos experimentalmente

Partindo do trabalho de Schonbrunn et al. [2013], foram extraídos resíduos e átomos do sítio de ligação da CDK2 que o autor destacou como importantes para a interação com ligantes. De maneira similar, no trabalho de Ho et al. [2009], foram extraídos resíduos e átomos do sítio ativo da ricina destacados pelos autores como relevantes na interação da ricina com subunidades 28S rRNA.

Figura 4.10. Grafos do grupo 1 da Ricina em que o maior padrão (3 vértices) para o suporte 0.1 está contido.

Com o intuito de validar os padrões encontrados pelo GReMLIN, foram comparados resíduos e átomos destes padrões com resíduos e átomos relevantes determinados experimentalmente para CDK [Schonbrunn et al., 2013] e Ricina [Ho et al., 2009].

4.5.1 Análise de resíduos importantes para CDK2

No maior padrão (5 vértices) dos dados da CDK2, foram encontrados os átomos LEU83:O (aceptor) e LEU83:N (doador) que interagem com ligantes. Ambos os átomos são relevantes na interação proteína-ligante da CDK2 e eles estão em uma região de articulação da proteína. Neste padrão, também foi possível encontrar HIS84:O (aceptor). Estes átomos estabelecem ligação de hidrogênio com os ligantes.

O grupo 3 possui um tipo de padrão com 2 vértices (um doador e um aceptor estabelecendo uma ligação de hidrogênio). Este padrão é encontrado em 73 dos 75 subgrafos. GLU81:O (aceptor), um resíduo de região de articulação da proteína, está presente em vários grafos do grupo 3 (3QWJ:A grafo 69, 3QX2:A grafo 78, 3QXO:A grafo 89, 3QZI:A grafo 113). Além disso, ASP86:N (doador) aparece em

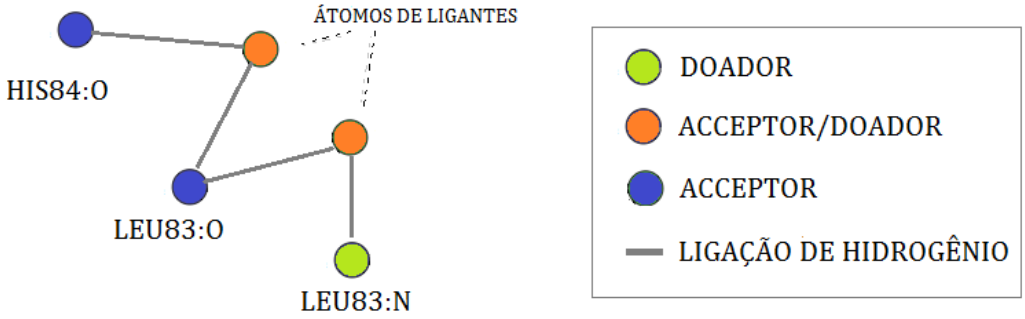


Figura 4.11. Padrão de 5 vértices onde estão os átomos LEU83:O, LEU83:N e HIS84:O.

vários grafos neste grupo (3RPV:A grafo 307, 3RPY:A grafo 312; 3S1H:A grafo 332; 3SQQ:A grafo 337). O átomo ASP1:145:OD1 (aceptor/negativo) aparece neste grupo, porém fora dos subgrafos frequentes.

O grupo 1 contém um padrão composto por dois vértices hidrofóbicos estabelecendo uma interação hidrofóbica. Este padrão é encontrado em 96 dos 157 grafos deste grupo. Átomos de interações hidrofóbicas relevantes da CDK2 aparecem como uma estrutura frequente neste grupo, por exemplo: LYS33:CB, LYS33:CD, LYS33:CE, LYS33:CG, LYS89:CB, LYS89:CE, ASP86:CB, ASP145:CG, ASP145:CB, PHE80:CB. Contudo, outros átomos importantes neste grupo, tais como ASP86:OD1 e ASP86:OD2 (ambos aceptor/negativo); LYS89:NZ (dador/positivo); PHE82:CZ, e PHE82:CE2 (ambos aromático/hidrofóbico); PHE80:CG, PHE80:CD2, PHE80:CE2, e PHE80:CZ (todos aromático/hidrofóbico), não estão contidos em um padrão (não estão em subgrafos frequentes). LYS33:NZ (doador/positivo) aparece 14 vezes no grupo 1, mas não está em uma estrutura frequente devido o seu tipo. O resíduo GLN85 é envolvido em uma interação com ligante mediada por água. Consequentemente, GLN85 não está em um subgrafo frequente já que GReMLIN considera somente interações diretas entre proteína-ligante. Esses resultados são sumarizados na Tabela 4.3.

4.5.2 Análise de resíduos importantes para Ricina

O maior padrão encontrado nos dados da ricina (8 vértices) apresentam os átomos CD1, CG, CE1, CZ, CE2, todos eles com o rótulo aromático/hidrofóbico de TYR80. No padrão de 5 vértices do grupo 9, foi encontrado CD2, CE2 e CG de TYR123. No padrão do grupo 5 (2 vértices aromático/hidrofóbico da proteína ligados a um vértice hidrofóbico do ligante) foi encontrado TYR80:CD2. É importante ressaltar que esses

Tabela 4.3. Resíduos do sítio de ligação da CDK2 interagindo com os dois mais potentes inibidores análogos à sulfonamida.

Resíduo	Átomo	GReMLIN
ASP145	CB	✓
	CG	✓
	OD1	•
LYS33	CB	✓
	CD	✓
	CE	✓
	CG	✓
	NZ	•
ASP86	N	✓
	CB	✓
	OD1	•
	OD2	•
GLN85	-	×
HIS84	O	✓

Resíduo	Átomo	GReMLIN
LYS89	CB	✓
	CE	✓
	NZ	•
LEU83	N	✓
	O	✓
PHE82	CE2	•
	CZ	•
GLU81	O	✓
PHE80	CB	✓
	CG	•
	CD2	•
	CE2	•
	CZ	•

✓ Resíduos encontrados nos padrões; • Encontrados fora dos padrões; × Não encontrados.

resíduos podem aparecer em outros padrões de outros grupos (por exemplo, TYR123 no grupo 7). Estes são alguns exemplos de padrões do GReMLIN envolvidos em interações ricina-ligante de acordo com resultados experimentais. Estes resíduos são apresentados na Tabela 4.4.

O resíduo VAL81:N (doador) foi encontrado no grupo 5. No grupo 1, temos VAL81:O (aceptor), GLY121:O (aceptor), ARG180 NH1 e NH2 (doador/positivo), ASP96:OD1 e OD2 (aceptor/negativo), ASP100:OD1 e OD2 (aceptor/negativo), ASN78:ND2(doador), GLU177:OE2 (aceptor/negativo), e TYR123:N (doador). Todos esses átomos mencionados aparecem em estruturas não frequentes. O grupo 1 é grande e heterogêneo, envolvendo vários tipos de átomos, o que torna difícil encontrar subestruturas frequentes com valores altos de suporte. Um caminho para lidar com essa questão é considerar outros valores de suporte para grupos heterogêneos, onde fosse permitido obter estruturas com pouca ocorrência, porém relevantes.

Em Ho et al. [2009], os autores destacaram TYR80 e TYR123 como resíduos relevantes que estabelecem interações π-*staking* (aromáticas). É importante notar que GReMLIN foi capaz de encontrar todos os átomos TYR80 e TYR123 em subgrafos frequentes (padrões), menos TYR123:N, que aparece nos dados resultantes do

Tabela 4.4. Resíduos do sítio ativo da cadeia A da Ricina interagindo com inibidores análogos ao estado cíclico de transição.

Resíduo	Átomo	GReMLIN
GLY121	O	•
ARG180	NH1	•
	NH2	•
ASP96	OD1	•
	OD2	•
ASP100	OD2	•
ASP75	OD2	•
ASN78	ND2	•
GLU208*	-	×
GLU177	OE2	✓
ARG134*	-	×

Resíduo	Átomo	GReMLIN
TYR80	CD1	✓
	CD2	✓
	CE1	✓
	CE2	✓
	CG	✓
	CZ	✓
TYR123	N	•
	CD2	✓
	CE2	✓
	CG	✓
VAL81	N	•
	O	•

✓ Resíduos encontrados nos padrões; • Encontrados fora dos padrões; × Não encontrados.

GReMLIN, mas não contido em uma subestrutura frequente. GLU208 e ARG134 não interagem diretamente com o ligante [Ho et al., 2009], logo o GReMLIN, devido a sua modelagem, é incapaz de considerá-los.

Capítulo 5

CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho propôs o GReMLIN, uma estratégia para tentar inferir padrões de interação em interfaces proteína-ligante, utilizando mineração de subgrafo frequente. A estratégia modela as interfaces proteína-ligante como grafos bipartidos onde os vértices representam átomos da proteína ou ligante e arestas representam as interações entre eles. Os vértices são rotulados de acordo com suas propriedades físico-químicas e as arestas são rotuladas de acordo com o tipo dos átomos e um critério de distância. Em sequência, uma análise de agrupamentos é conduzida sobre os grafos criados e um algoritmo para mineração de subgrafo frequente é executado sobre estes mesmos grafos para encontrar subestruturas conservadas na interface proteína-ligante de uma família de proteínas.

Os resultados mostram que GReMLIN é capaz de encontrar subestruturas frequentes nas interfaces proteína-ligante. Além disso, quando comparadas as interações experimentalmente determinadas, nossa estratégia *in silico* proposta foi capaz de encontrar vários resíduos relevantes no sítio de ligação de inibidor para CDK2 e resíduos de sítio ativo para Ricina. Este trabalho resultou em uma publicação na conferência BIBE 2016 (IIEEE 16th International Conference on Bioinformatics and Bioengineering), com o prêmio de trabalho de destaque. A ferramenta para visualização dos resultados gerou um artigo submetido no Journal of Bioinformatics and Computational Biology. Ambos os textos estão disponíveis nos Apêndices A e B respectivamente.

Como trabalhos futuros, pretendemos melhorar a estratégia de agrupamento, já que os resultados da mineração de subgrafo frequente são altamente dependentes da qualidade dos grupos. Por exemplo, o grupo 1 da ricina possui diversos conjuntos

de grafos contendo tipos e topologias bem diversificados que dificultou a busca de padrões de alta frequência. Além disso, é preciso considerar moléculas de água para calcular interações mediadas por água, assim como, incluir em nossa modelagem os átomos de hidrogênio existentes nas interações intermoleculares.

Com o intuito de complementar este trabalho, pretendemos utilizar os padrões encontrados pelo GReMLIN para prever potenciais ligantes pra uma proteína alvo. Usando como estudo de caso os dados da Ricina e CDK2 humana, será feita uma busca em bases de dados publicamente disponíveis na internet por ligantes que contenham os padrões encontrados pelo GReMLIN. Para validar os ligantes, estes serão submetidos a etapas experimentais para verificar a afinidade com sua respectiva proteína alvo.

Referências Bibliográficas

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

Buxbaum, E. (2015). *Fundamentals of protein structure and function*, volume 31. Springer.

Campos, A. R.; Gao, Z.; Blaber, M. G.; Huang, R.; Schatz, G. C.; Van Duyne, R. P. & Haynes, C. L. (2016). Surface-enhanced raman spectroscopy detection of ricin b chain in human blood. *The Journal of Physical Chemistry C*, 120(37):20961--20969.

Chen, H. Y.; Foo, L. Y. & Loke, W. K. (2015). Ricin and abrin: A comprehensive review of their toxicity, diagnosis, and treatment. *Toxinology*.

Cheng, T.; Li, Q.; Zhou, Z.; Wang, Y. & Bryant, S. H. (2012). Structure-based virtual screening for drug discovery: a problem-centric review. *The AAPS journal*, 14(1):133--141.

Claude Sammut, G. I. W. e. (2010). *Encyclopedia of Machine Learning*. Springer US, 1 edição.

Cordella, L. P.; Foggia, P.; Sansone, C. & Vento, M. (2004). A (sub) graph isomorphism algorithm for matching large graphs. *IEEE transactions on pattern analysis and machine intelligence*, 26(10):1367--1372.

Cormen, T.; Leiserson, C.; Rivest, R. & Clifford, S. (2010). *Introduction to Algorithms. 3rd Edn. Vol. 1*. Cambridge, MA: MIT Press.

Elseidy, M.; Abdelhamid, E.; Skiadopoulos, S. & Kalnis, P. (2014). Grami: Frequent subgraph and pattern mining in a single large graph. *Proceedings of the VLDB Endowment*, 7(7):517--528.

Fromm, H. J. & Hargrove, M. (2012). *Essentials of biochemistry*. Springer Science & Business Media.

Fuller, J. C.; Martinez, M.; Henrich, S.; Stank, A.; Richter, S. & Wade, R. C. (2015). Ligdig: a web server for querying ligand–protein interactions. *Bioinformatics*, 31(7):1147--1149.

Garbuzynskiy, S. O.; Melnik, B. S.; Lobanov, M. Y.; Finkelstein, A. V. & Galzitskaya, O. V. (2005). Comparison of x-ray and nmr structures: Is there a systematic difference in residue contacts between x-ray-and nmr-resolved protein structures? *Proteins: Structure, Function, and Bioinformatics*, 60(1):139--147.

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

Gu, J. & Bourne, P. E. (2009). *Structural bioinformatics*, volume 44. John Wiley & Sons.

Guedes, I. A.; de Magalhães, C. S. & Dardenne, L. E. (2014). Receptor–ligand molecular docking. *Biophysical Reviews*, 6(1):75--87.

Han, J.; Pei, J. & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Ho, M.-C.; Sturm, M. B.; Almo, S. C. & Schramm, V. L. (2009). Transition state analogues in structures of ricin and saporin ribosome-inactivating proteins. *Proceedings of the National Academy of Sciences*, 106(48):20276--20281.

HOFFMANN, L.; Dantas, A.; de Medeiros, E. & Soares, L. (2007). Ricina: um impasse para utilização da torta de mamona e suas aplicações. *Documentos*.

Huan, J.; Wang, W. & Prins, J. (2003). Efficient mining of frequent subgraphs in the presence of isomorphism. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pp. 549--552. IEEE.

Huan, J.; Wang, W.; Prins, J. & Yang, J. (2004). Spin: mining maximal frequent subgraphs from graph databases. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 581--586. ACM.

Huang, S.-Y.; Grinter, S. Z. & Zou, X. (2010). Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. *Physical Chemistry Chemical Physics*, 12(40):12899--12908.

Jiang, C.; Coenen, F. & Zito, M. (2013). A survey of frequent subgraph mining algorithms. *The Knowledge Engineering Review*, 28(01):75--105.

Johnson, L. N. (2009). Protein kinase inhibitors: contributions from structure to clinical compounds. *Quarterly reviews of biophysics*, 42(01):1--40.

Kaufman, L. & Rousseeuw, P. J. (1990). Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, pp. 68--125.

Ketkar, N. S.; Holder, L. B. & Cook, D. J. (2005). Subdue: Compression-based frequent pattern discovery in graph data. In *Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations*, pp. 71--76. ACM.

Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A. et al. (2015). Pubchem substance and compound databases. *Nucleic acids research*, p. gkv951.

Klein, R. & Lee, D. (2014). Voronoi diagrams an delaunay triangulations. *Encycl Algorithm*, pp. 1--5.

Klema, V. & Laub, A. (1980). The singular value decomposition: Its computation and some applications. *IEEE Transactions on automatic control*, 25(2):164--176.

Konc, J. & Janežič, D. (2014). Probis-ligands: a web server for prediction of ligands by examination of protein binding sites. *Nucleic acids research*, 42(W1):W215--W220.

Koyutürk, M.; Grama, A. & Szpankowski, W. (2004). An efficient algorithm for detecting frequent subgraphs in biological networks. *Bioinformatics*, 20(suppl 1):i200--i207.

Laskowski, R. A. & Swindells, M. B. (2011). Ligplot+: multiple ligand–protein interaction diagrams for drug discovery.

Lee, J.; Han, W.-S.; Kasperovics, R. & Lee, J.-H. (2012). An in-depth comparison of subgraph isomorphism algorithms in graph databases. In *Proceedings of the VLDB Endowment*, volume 6, pp. 133--144. VLDB Endowment.

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

Mannhold, R.; Kubinyi, H.; Folkers, G.; Böhm, H.-J. & Schneider, G. (2006). *Protein-ligand interactions: from molecular recognition to drug design*, volume 19. John Wiley & Sons.

Nelson, D. L.; Lehninger, A. L. & Cox, M. M. (2014). *PrincÃpios da BioquÃmica 6.ed.* Artmed.

Pang-Ning, T.; Steinbach, M.; Kumar, V. et al. (2006). Introduction to data mining. In *Library of congress*, volume 74.

Pires, D. E.; de Melo-Minardi, R. C.; da Silveira, C. H.; Campos, F. F. & Meira, W. (2013). acsm: noise-free graph-based signatures to large-scale receptor-based ligand prediction. *Bioinformatics*, 29(7):855--861.

Rokach, L. (2009). A survey of clustering algorithms. In *Data mining and knowledge discovery handbook*, pp. 269--298. Springer.

Saidi, R.; Maddouri, M. & Nguifo, E. M. (2009). Comparing graph-based representations of protein for mining purposes. In *Proceedings of the KDD-09 Workshop on Statistical and Relational Learning in Bioinformatics*, pp. 35--38. ACM.

Salentin, S.; Schreiber, S.; Haupt, V. J.; Adasme, M. F. & Schroeder, M. (2015). Plip: fully automated protein–ligand interaction profiler. *Nucleic acids research*, 43(W1):W443--W447.

Sammut, C. & Webb, G. I. (2011). *Encyclopedia of machine learning*. Springer Science & Business Media.

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.

Sedgewick, R. & Flajolet, P. (2013). *An introduction to the analysis of algorithms*. Addison-Wesley.

Shelokar, P.; Quirin, A. & Cordón, Ó. (2013). A multiobjective evolutionary programming framework for graph-based data mining. *Information Sciences*, 237:118--136.

Shen, R. & Guda, C. (2014). Applied graph-mining algorithms to study biomolecular interaction networks. *BioMed research international*, 2014.

Silva, B. V.; Horta, B. A.; Alencastro, R. B. d. & Pinto, A. C. (2009). Proteínas quinases: características estruturais e inibidores químicos. *Química Nova*, 32(2):453--462.

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

Sterling, T. & Irwin, J. J. (2015). Zinc 15-ligand discovery for everyone.

Tan, P.-N. et al. (2006). *Introduction to data mining*. Pearson Education India.

Thomas, L. T.; Valluri, S. R. & Karlapalem, K. (2010). Margin: Maximal frequent subgraph mining. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(3):10.

Vishveshwara, S.; Brinda, K. & Kannan, N. (2002). Protein structure: insights from graph theory. *Journal of Theoretical and Computational Chemistry*, 1(01):187--211.

Wallace, A. C.; Laskowski, R. A. & Thornton, J. M. (1995). Ligplot: a program to generate schematic diagrams of protein-ligand interactions. *Protein engineering*, 8(2):127--134.

Wei, D.; Xu, Q.; Zhao, T. & Dai, H. (2015). *Advance in Structural Bioinformatics*. Springer.

Williams, M. A. & Daviter, T. (2013). *Protein-ligand interactions*. Springer.

Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z. & Woolsey, J. (2006). Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34(suppl 1):D668--D672.

Xiong, J. (2006). *Essential bioinformatics*. Cambridge University Press.

Yan, X. & Han, J. (2002a). gspan: Graph-based substructure pattern mining. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pp. 721--724. IEEE.

Yan, X. & Han, J. (2002b). gspan: Graph-based substructure pattern mining. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pp. 721--724. IEEE.

Zaki, M. J.; Meira Jr, W. & Meira, W. (2014). *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press.

Apêndice A

Artigo Publicado

Santana, Charles A., Fabio R. Cerqueira, Carlos H. da Silveira, Alexandre V. Fassio, Raquel C. de Melo-Minardi, and Sabrina de A. Silveira. "**GReMLIN: A graph mining strategy to infer protein-ligand interaction patterns**". In IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE 2016), pp. 28-35. IEEE, 2016. **Distinguished Student Paper Award.**

GReMLIN: A graph mining strategy to infer protein-ligand interaction patterns

Charles A. Santana*, Fabio R. Cerqueira*, Carlos H. da Silveira†, Alexandre V. Fassio‡,
Raquel C. de Melo-Minardi‡ and Sabrina de A. Silveira*
*Computer Science Department
Universidade Federal de Viçosa, Viçosa, Brazil 36570–900
Email: {charles.santana, fabio.cerqueira, sabrina}@ufv.br
†Advanced Campus at Itabira
Universidade Federal de Itajubá, Itabira, Brazil
Email: carlos.silveira@unifei.edu.br
‡Universidade Federal de Minas Gerais, Belo Horizonte, Brazil 31270–901
Email: {alexandre, raquelcm}@dcc.ufmg.br

***Abstract*—Interactions between proteins and ligands are relevant in many biological processes. In the last years, such interactions have gained even more attention as the comprehension of protein-ligand molecular recognition is an important step to ligand prediction, target identificantion, and drug design, among others. This article presents GReMLIN (Graph Mining strategy to infer protein-Ligand INteraction patterns), a strategy to search for conserved protein-ligand interactions in a set of related proteins, based on frequent subgraph mining, that is able to perceive structural arrangements relevant for protein-ligand interaction. When compared to experimentally determined interactions, our *in silico* strategy was able to find many of relevant binding site residues/atoms for CDK2 and active site residues/atoms for Ricin.**

Availability: www.dcc.ufmg.br/~sabrinas/visgremlin

***Keywords*-graph; interaction; protein; ligand; pattern**

I. Introduction

Ligand-receptor interactions are pivotal in a variety of biological processes in living systems. Signal transmission via molecular complementarity, regulation of biological function, and conformational change in protein through cooperativity in ligand binding are some of these processes [1], to name a few. However, a key factor that has been atracted much attention is that understanding the ligand-receptor recognition process is a major step towards ligand prediction, target identification, lead discovery, and drug design [2].

Over 119.000[1] entries of structural data are currently available in Protein Data Bank (PDB) [3]. Also, a comprehensive catalogue of compounds including chemical and structural information are provided in public databases as PubChem [4], Zinc [5], and DrugBank [6]. These resources made possible to study protein-compound interactions from many perspectives, as protein-ligand interactions (PLI) networks, ligand based virtual screening (*vs*) through fingerprints, structure based *vs* through docking, among others.

[1]http://www.rcsb.org/pdb/statistics/holdings.do

In this work, we investigate whether it is possible to find patterns that characterize protein-ligand interactions specific for a set of related proteins. Such patterns can be seen as inteligible key factors to support the understanding of molecular recognition process. Moreover, if such patterns exist, we believe they can represent a step towards the prediction of protein-ligand interaction. Here, we consider as ligands small non-protein molecules.

Many methods have been proposed to study protein-ligand interactions. We do not aim at covering an exhaustive list of such methods. Here, we briefly review some representative cases. In [7], the authors proposed LigPlot+, a software tool to generate 2D diagrams of ligand-protein interactions from 3D coordinates which allows to superimpose such 2D diagrams from different protein-ligand structures and highlights conserved interactions for residues in corresponding 3D positions. Although it is not a method whose the ultimate goal is detecting patterns in a set of protein-ligand structures, the tool is able to detect conserved interactions, but for a small set of structures, as its interface becomes cluttered for larger sets.

In the context of machine learning (ML), the proposed methods to find protein-ligand interactions can be divided in three types, according to [8]. The first one is *feature vector-based*, which takes input instances (e.g., drug-target interactions) represented by feature vectors (e.g., a set of chemical descriptors for compounds and proteins). This kind of input can feed a variety of machine learning methods, such as SVM, KNN, and Naive Bayes which were used, for instace, in the works [9], [2].

The second type of ML methods is *similarity-based*, which takes as input a similarity matrix, where the position i, j represents the similarity between element i and j (e.g., a similarity matrix for compounds can be computed considering their chemical structures. For target proteins, a similarity based matrix can be computed by sequence alignment). Representative examples of this thechnique are Gaussian interaction profile [10] and kernelized Bayesian

matrix factorization [11].

The last type of ML involves approaches that extract implicit co-occurrent compound-protein relations through text mining techniques from many kinds of information, as biomedical literature [12], [13]. Nonetheless, these relations might not represent real compound-protein interactions.

Although there are many methods to study and predict protein-ligand interactions, few of them are able to identify which features imply in protein-ligand affinity, i.e., which are the key factors to perform the prediction.

This article proposes GReMLIN (**Gr**aph **M**ining strat**e**gy to infer protein-**L**igand **IN**teraction patterns), a scalable graph based strategy which models protein-ligand interfaces as bipartite graphs where nodes depict protein or ligand atoms and edges depict the interactions among them. Nodes are labeled with physicochemical properties of atoms and edges are labeled according to the atoms properties and a distance criteria. A clustering analysis is performed on the computed graphs and we conduct a frequent subgraph mining on each cluster to infer relevant patterns in protein-ligand interface. Frequent subgraphs computed by GReMLIN express structural arrangements relevant for protein-ligand interaction and molecular recognition.

We provide an online tool to visualize and explore the interaction patterns obtained for test datasets as well as the result of clustering analysis and frequent subgraph mining. Also, we provide a 3D representation of all protein-ligand interaction graphs in a molecule viewer. All the source codes and datasets are available in the GReMLIN website.

II. Methods

In this section, we describe the datasets used, the modeling of protein structures as graphs at the atomic level, the graph mining task proposed to infer protein-ligand patterns, and the evaluation strategy.

A. Data

We collected structural data of proteins in complex with ligands from PDB to validate our method and show its applicability in a real world scenario. We have two datasets, Ciclin dependent kinases (*CDK*) 2 and *Ricin*, both with biological relevance but different features.

CDK2: 73 PDB entries with identical sequences coupled with different ligands. CDKs are a family of serine/threonine kinases involved in cell cycle progression and transcription. As its deregulation has been associated with a number of medical conditions, such as cancer and neurodegenerative disorders, they become important targets in drug discovery [14], [15]. This dataset was derived from [15], in which relevant residues for PLI were experimentally determined. We intend to compare patterns obtained with our strategy to those experimentally determined to examine if GReMLIN is able to find important patterns in PLI. The CDK2 dataset entries from PDB are provided in the GReMLIN website.

Ricin: 29 PDB entries, which share sequence identity greater than or equal to 50% with ricin template 2AAI.A. Ricin has a heterodimeric structure consisting of an A-chain linked by a disulfide bond to a B-chain. B-chain is responsible for carring A-chain into the cell where it catalyzes adenosine depurination of 28S rRNA to inhibit protein synthesis causing cell death. Due to high toxicity, stability, ease of production, and availability of castor beans, ricin is considered a potential agent of chemical/biological warfare of interest to terrorists [16]. However, accordding to [17], a comprehensive understanding of the pathways exploited by ricin is missing. We believe our strategy can support broader understanding of ricin interactions with its ligands. This dataset is used to compare PLI patterns experimentally obtained in [18] with GReMLIN patterns. The *Ricin* dataset entries from PDB are provided in the GReMLIN web site.

Among the test datasets, we consider that Ricin is more difficult and realistic. Ricin structural data is scarce [18], which is a challenge for data mining algorithms. Also, in this dataset, we have a set of different proteins interacting with different ligands, which is a more realistic scenario because when a domain specialist has a dataset of interest, it can be composed of similar (not identical) proteins.

B. Modeling

Interaction computation

From the structures obtained in PDB, we used those containing one or more ligands. Ligands with less than 6 atoms were considered as cristalographic artifacts and were then removed. In this text, we use contacts and interactions interchangeably. We computed the contacts at the atomic level by using Voronoi diagram and Delaunay triangulation [19], which comprise a geometric-based and cutoff independent approach in which edges represent interactions between atoms that are probably in direct contact, avoiding occlusion.

Graph generation

We parsed the whole set of Delaunay edges to select only those in the protein-ligand interface, which means that only edges that connect a protein atom with a ligand atom were kept. We can formally state the set of interactios between atoms as a bipartite graph $G = (P, L, E)$, whose vertices can be divided in two disjoint sets, P and L, such that every edge in E connects a vertex in P to a vertex in L. Thus, the set of vertices P represent protein atoms, the set of vertices L represent ligand atoms, and the set of edges E represent the interactions between a protein atom and a ligand atom.

As different regions from a protein can potentially establish interactions with different ligands, we computed the connected components for each PDB entry, so that each entry can result in one or more connected graphs.

A graph is connected when there is a path between every pair of vertices. In this work, each connected component is considered a graph for the graph mining purpose and each graph can represent protein-ligand interactions for one or more ligands.

For instance, consider a protein node (atom) from the Ricin dataset which establishes 2 interactions. In a PDB entry X, this protein node (atom) interacts with 2 atoms, each one from a different ligand, so that these ligands are part of the same connected component in our modeling. In a protein entry Y, this protein node interacts with 2 atoms from the same ligand. If we consider a graph for each protein-ligand pair of a PDB entry, when we mine frequent interactions for this protein dataset, we will not detect the 2 interactions established for this protein node as a pattern. Therefore we believe that considering each connected component as a graph is more appropriate than considering each protein-ligand pair as a graph for a specific PDB entry.

Graph labeling

For each graph, we labeled nodes and edges according to physicochemical properties and distance criteria. Vertices from proteins were labeled as acceptor, aromatic, donor, hydrophobic, negative (anion) or positive (cation) according to our previous work [20], which, in turn, was derived from [21]. Vertices from ligands were labeled with the same physicochemical properties, computed by Pmapper from Chemaxon[2] at pH 7 (Pmapper 5.3.8, 2010). Edges were labeled based on its vertices physicochemical properties and a distance criteria according to [22]. Table I provides the distance criteria and atom types for each interaction.

Table I
DISTANCE CRITERIA (IN Å) CONSIDERED TO COMPUTE INTERACTIONS.

Interaction type	Atom types	Min. distance	Max. distance
Aromatic stacking	2 aromatic atoms	1.5	3.5
Hydrogen bond	1 acceptor and 1 donor atom	2.0	3.0
Hydrophobic	2 hydrophobic atoms	2.0	3.8
Repulsive	2 atoms with the same charge	2.0	6.0
Salt bridges	2 atoms with opposite charge	2.0	6.0

C. Experiments

We conducted two experiments to search for patterns in protein-ligand interactions for our datasets. After the process of computing graphs, each dataset, CDK2 and Ricin, resulted in a set of graphs. For each dataset, we first performed a clustering analysis to characterize our datasets and to facilitate the understanding of similarities and differences between interactions that compose the protein-ligand interface. Next, we searched for frequent sugraphs in each cluster through a frequent subgraph mining algorithm. Our hypothesis is that if a subgraph is frequent in a cluster, then it represents a relevant pattern in protein-ligand interactions in that group.

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

Clustering analysis

Graph dataset summary: To summarize our graph dataset, we propose a counting matrix in which graphs are represented based on the labels of vertices in the end of edges. In this matrix, each row represents a graph and each column represents a pair of node labels in the end of an edge. Suppose, for instance, that we have node labes A (aromatic), B (acceptor), and C (donor) as well as graphs G1, G2, and G3 in a certain dataset. The counting matrix for this example dataset is provided in Table II. If position (i, j) has value v, it means that Graph G_i has v edges whose ends are the labels represented by column j. For instance, the position $(1, 5)$ in Table II is 3, which means that the graph $G1$ has 3 edges whose ends are labels B (acceptor) and C (donor). For our ricin dataset, we have 181 rows and 23 columns. For CDK dataset, we have 339 rows and 16 columns.

Table II
EXAMPLE OF A COUNTING MATRIX

Protein chain	AA	AB	AC	BB	BC	CC
G1	2	0	1	0	3	0
G2	0	0	0	0	1	0
G3	0	0	1	0	0	4

Noise reduction: After computing the counting matrix, which we named A, a dimensionality and noise reduction step is applied using Singular Value Decompositon (SVD). To compress the data used in the clustering analysis, reducing the number of columns and noise in matrix A, yet keeping relevant semantic relationships among the terms, the matrix A can be approximated by matrix A_d (with rank d, where d is less than the rank of A) using: $A_d = U_d \Sigma_d V_d^T$.

To achieve A_d, the first d singular values of A and their singular vectors were taken, and, thus, the resulting matrix has d features: $A_d = U_d \Sigma_d V_d^T = U_d(\Sigma_d V_d^T) = U_d(E_d)$. A_d can be approximated by matrix E_d, which is: $E_d = \Sigma_d V_d^T$ [23]. Here, the counting matrix A was approximated by E_d.

SVD was used in a similar manner in [24], [25]. According to [26], the choice of d is empirical. Thereupon, approximations with all possible values of d were computed, and the matrix that led to the best clustering was selected.

The counting matrices' dimensions are not critical for a clustering analysis. Nonetheless, as such matrices are sparse, we can benefit from SVD noise reduction.

Clustering: This counting table is used to perform a clustering analysis of our datasets, which aims to generate clusters

or groups of instances (graphs in our case), so that objects within a cluster have high similarity among each other, but are very dissimilar to objects in other clusters. To group our graph data, we used the partitioning algorithm k-medoids to cluster graphs in k partitions, with $1 < k < n$, where n is the number of graphs in a dataset. We do not tried $k = 1$ or $k = n$ as they are the trivial partitions, where all instances are in the same group or each instance is alone in a group, respectively. The implementation of k-medoids $pamk$ in the fpc package from the R[3] software version 3.0.2 was used in this experiment. This implementation of k-medoids performs a partitioning around medoids with the number of clusters estimated by optimum *average silhouette width* (asw).

The reduced matrices from SVD are given as input to k-medoids. For each reduced matrix, we varied the number of groups, k, in k-medoids to select the best clustering.

Frequent subgraph mining

We conducted a *frequent subgraph mining* (FSM) experiment to search for frequent patterns in each resulting cluster from best clusterings (we have 2 best cluterings, one for Ricin and one for CDK). In this work, by frequent patterns we mean frequent subgraphs.

FSM was performed with the gSpan algorithm [27]. According to the gSpan authors, given a graph dataset $D = \{G_0, G_1, ..., G_n\}$, $support(g)$ denotes the number of graphs in D which have g as a subgraph. Then, the FSM is to find any subgraph g whose $support(g) \geq minSup$ (a minimum support threshold).

For each cluster, we executed gSpan with $support$ varying from 0.1 to 1.0. As support increases, we obtain subgraphs that are in a high fraction in the graph input dataset, but the number of total subgraphs tend to decrease. Also, as we increase the $support$, the resulting subgraphs tend to be small, which is expected as it is difficult to find large patterns present in the whole graph input dataset.

According to [28], in biological networks, maximal frequent subgraphs are deemed to be the most interesting ones. Also, the resulting subgraphs from gSpan can be structurally repetitive, as a frequent subgraph can present other frequent subgraphs within it [29]. Thus, we filtered only the maximal subgraphs for the subsequent analyses. A subgraph g is maximal frequent if it satisfies the following conditions: (i) g must be frequent and (ii) there must be no frequent super graph of g [30]. Extracting maximal frequent subgraphs considerably reduces the number of subgraphs found.

D. Evaluation strategy

Next, we explain the evaluation strategy employed in each experiment.

[3]https://www.R-project.org/

Clustering analysis

The process we detail below is the same for the Ricin and CDK datasets. We performed a clustering analysis with all matrices, which means a matrix for each number of dimensions d resultant from SVD. For each matrix dimension d, we conducted a clustering analysis to select the matrix that results in the best clustering.

For each matrix, we applied the k-medoids algorithm varying the number of groups k from $k = 2$ up to $k = n-1$, where n is the number of graphs in the dataset. The quality of clusterings was assessed by comparing their asw, i.e. the choice of k in k-medoids was based on the highest value for asw. A high asw value for a clustering means that the algorithm discovered a very strong clustering structure [31]. On the other hand, when the algorithm leads to a poor clustering, the overall asw tends to be very low. Hence silhouette plots and averages may be used to determine the natural number of clusters in a dataset.

Frequent subgraph mining

In order to select an appropriate $support$ for an FSM experiment, we generated a table that provides, for each cluster and for each $support$ value from 0.1 to 1.0, the frequency and size (number of vertices) of resulting frequent subgraphs. In this table, we use a heatmap in which the color is a pre-attentive attribute that encodes the frequency of subgraphs (this table is available in our visualization tool). The choice for $support$ was empirical and varying its value provide us with an overview of the size and frequency of patterns. There is a compromise between large and frequent subgraphs. The larger the size, the smaller the frequency. In Section III, we further discuss our $support$ choice.

The FSM algorithm outputs the information of what are the frequent subgraphs in a dataset of graphs and in which input graphs they appear. However, it is not enough to point out in the input dataset which are the nodes and edges that correspond to a specific subgraph. Therefore, after computing frequent subgraphs, we need to map them to the input graphs, a task that was performed using VF2 [32], a subgraph isomorfism algorithm.

Protein-ligand patterns obtained by GReMLIN are compared to patterns experimentally determined for Ricin and CDK to check if our strategy is able to find patterns experimentally determined.

III. Results and discussion

A. Clustering analysis

The best clustering for CDK2 dataset was obtained using the resulting matrix from SVD with the best asw (1.0). All resulting matrices from SVD with d variyng from 1 to 15 were tested, as well as values for k varying from 2 to 338.

As a result, and the best silhouette coefficient was reached with $d = 1$ and $k = 5$.

For Ricin dataset, the experiment configuration that led to the best asw (0.93) was $d = 1$ for the matrix resulting from SVD along with a number of groups $k = 9$ in k-medoids. All possible values for d and k were tested in a similar manner to which the CDK2 dataset was processed.

B. Analysis of frequent patterns

We are interested in the most frequent subgraphs, which are the common frequent substructures in protein-ligand interactions. However, as we increase the $support$ value, we lose some patterns. Hence, there is a compromise between the $support$ value and the number of patterns found. Next, we discuss some selected patterns from CDK2 and Ricin.

CDK2

For the CDK2 dataset, we chose the $support$ value 0.6, based on our *Simple table* in *Graph patterns table*. We observe in this table that from $support$ 0.6 to 0.7 we lose one pattern of size 2 in *Group 1*. As a consequence, so we decided to choose $support$ 0.6.

We briefly discuss some interesting patterns found with GReMLIN, starting from the largest pattern from the Ricin dataset (Figure 1). This pattern has 5 nodes and was found in *Group 5* in graph 8 (3QQH:A, ligand X0A) and graph 10 (3QQJ:A, ligand: X11). For both graphs, the nodes from protein are HIS84:O (acceptor), LEU83:O (acceptor), and LEU83:N (donor), while the ligand nodes differ for the 2 graphs. In graph 8, the ligand nodes are X0A:303:N04 (acceptor/donor) and X0A:303:O21 (acceptor/donor), while in graph 10, ligand nodes are X11:300:N8 and X11:300:O15 (acceptor/donor). In Figure 2, we highlight the atoms from these ligands which interact with CDK2. We observe that they have a part of their structures in common, which allows them to interact with the same protein structure.

It indicates that searching for frequent substructures in protein-ligand complexes interface can reval some relevant patterns which are determinant for protein-ligand interaction.

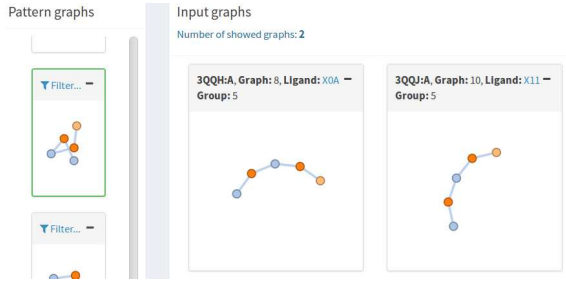
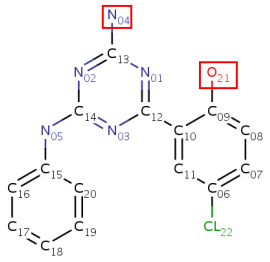
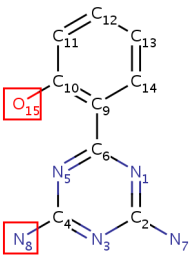


Figure 1. Pattern with 5 nodes from group 5 for the CDK2 dataset.

Groups 1, 2, and 3 have subgraphs with 2 nodes and 1 interaction. *Group 3* has 75 graphs and it is mainly composed of a pair of acceptor and donor nodes linked by a hydrogen bond edge.



(a) Ligand X0A atoms N04 and O21



(b) Ligand X11 atoms N8 and O15

Figure 2. The largest pattern from CDK2 (5 nodes) was found in group 5, in graphs 8 (3QQH:A) and 9 (3QQJ:A). We highlight the ligand atoms from these graphs that interact with CDK2.

Group 2 (101 graphs) is mainly composed of a pair of acceptor and acceptor/donnor atoms and, consequently, hydrogen bond interactions. In general, ligand atoms are labeled as acceptor/donnor and protein atoms are labeled as acceptor. Also, protein atoms are from LEU83 and GLU81 in most part of the graphs. The exceptions are the protein atom ASP132:OD1 in 3RPV:A (graph 309) labeled as acceptor/donor and the protein atom ASP86:OD2 from 3QZF:A (graph 97) labeled as acceptor/negative.

Group 1 exhibts primarily hydrophobic atoms and interactions (dark blue in our prototype tool). It has also some hydrophobic/aromatic (green) nodes as well as donor, donor/positive, acceptor/donor and acceptor/aromatic.

Ricin

The $support$ chosen for ricin dataset was 0.6 because, observing the *Simple table* from *Graph patterns table*, we noted that from $support$ 0.6 to 0.7 we lost 4 types of subgraphs with 5 nodes for group 4. Considering that we want the largest common substructures for each group, we believe 0.6 is a reasonable choice for the $support$.

We succinctly discuss some interesting patterns from Ricin dataset starting from the largest ones. This dataset has 1 pattern of 8 nodes, which was found in *Group 6* (containing 3 graphs) in graphs 26 (1IL5:A, ligand DDP) and 101 (3RTI:B, ligand GAL). In graph 26, all protein atoms are from TYR80 and in graph 101 all protein atoms are from TRP37. Protein atoms from both graphs are labeled as aromatic/hydrofobic and ligand atoms as hydrophobic. Therefore, in these graphs there are only hydrophobic interactions. Figure 3 shows the pattern of 8 nodes and its 2 mappings in the graphs of *Group 6*. In Figure 4, we highligh the atoms of ligands involved in this pattern. Also, *Group 6* has 6 subgraphs with 7 nodes, 3 subgraphs with 6 nodes, and 1 subgraph with 5 nodes, all of them involving hydrophobic

atoms from the ligand side and aromatic/hydrophobic atoms from the protein side, resulting in hydrophobic interactions.

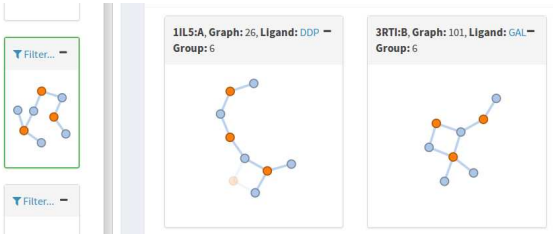
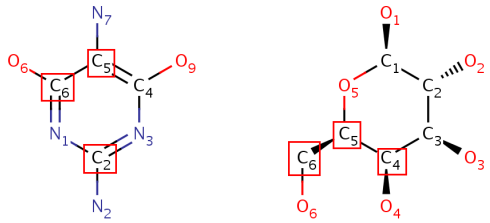


Figure 3. Pattern with 8 nodes from group 6 for the Ricin dataset.



(a) Ligand DDP atoms C2, C5, C6

(b) Ligand GAL atoms C4, C5, C6

Figure 4. The largest pattern from Ricin (8 nodes) was found in group 6, in graphs 26 (1IL5:A) and 101 (3RTI:B). We highlight the ligand atoms from these graphs that interact with Ricin.

Group 2 (contains 16 graphs) has 1 frequent subgraph with 2 nodes. The protein atoms are from TYR80, except for 1BR5:A that is from TYR123, labeled as aromatic/hydrophobic, while the ligand nodes are labeled as aromatic. Consequently, aromatic interactions are frequent in this group.

Although *Group 1* is the largest group in the Ricin dataset, with 130 graphs, it is a very diverse group and presents frequent subgraphs only for $support \geq 0.1$ (4 frequent subgraphs with 2 nodes and 1 frequent subgraph wiht 3 nodes). By inspecting this group on our visualization tool, we see by the predominant color of edges (yellow) that the majority of interactions are hydrogen bonds. However, their amount and shape (graph topology) are not conserved across the group. It is expected that the Ricin dataset is heterogeneous, as it is composed of PDB entries with diverse sequences ($identity \geq 50\%$) and varied ligands.

C. Comparison of GReMLIN with experimental patterns

From the work of Schonbrunn and colleagues [15], we extracted binding site residues and atoms that the authors highlight as important for CDK2 interaction with ligands. In that work, the authors described the development of highly potent and selective diaminothiazole inhibitors of CDK2 based on a single hit compound with weak inhibitory activity.

In a similar manner, from the work of Ho and colleagues [18], we extracted active site residues and atoms that the authors highlight as relevant in the interaction of ricin with subunit 28S rRNA. In that work, the authors co-crystalize ricin chain A with a transition state analogue inhibitor that mimics the sarcin-ricin recognition loop of the 28S rRNA.

We compare residues and atoms from frequent subgraphs and their groups computed by GReMLIN with relevant residues and atoms experimentally determined for CDK2 and Ricin in [15] and [18].

CDK2

In the largest pattern (5 nodes in *Group 5*) from CDK2, we found LEU83:O (acceptor) and LEU83:N (donor) as atoms that interact with ligands. Both atoms are relevant in CDK2 protein-ligand interaction and they are in the hinge region of this protein. In this pattern, we could also find HIS84:0 (acceptor). These atoms establish hydrogen bonds with ligands.

Group 3 has one type of pattern with 2 nodes (a donor and an acceptor node establishing a hydrogen bond). This pattern is found in 73 out of 75 patterns. GLU81:O (acceptor), a residue of hinge region, is found in many graphs from this group (i.e., 3QWJ:A graph 69; 3QX2:A graph 78; 3QXO:A graph 89; 3QZI:A graph 113). Also, ASP86:N (donor) appears in many graphs in this group (i.e., 3RPV:A graph 307; 3RPY:A graph 312; 3S1H:A graph 332; 3SQQ:A graph: 337). The atom ASP145:OD1 (acceptor/negative) appears in this group, but not in a pattern.

Group 1 contains one pattern composed of 2 hydrophobic nodes establishing a hydrophobic interaction. This pattern is found in 96 out of 157 graphs in this group. Relevant hydrophobic interacting atoms from CDK2 appear as a pattern in this group, for instance: LYS33:CB, LYS33:CD, LYS33:CE, LYS33:CG, LYS89:CB, LYS89:CE, ASP86:CB, ASP145:CG, ASP145:CB, PHE80:CB. However, other important atoms in this group, such as ASP86:OD1 and AS86P:OD2 (both acceptor/negative); LYS89:NZ (donor/positive); PHE82:CZ, and PHE82:CE2 (both aromatic/hydrophobic); PHE80:CG, PHE80:CD2, PHE 80:CE2, and PHE80:CZ (all aromatic/hydrophobic), are not a pattern (they are not in frequent subgraphs). LYS33:NZ (donor/positive) appears 14 times in this group, but is not in a pattern due to its type.

The residue GLN85 is involved in a water mediated interaction with ligand. Consequently, GLN85 is not in GReMLIN frequent subgraphs as our strategy considers only direct protein-ligand interactions. These results are summarized in Table III

Ricin

The largest pattern for Ricin (8 nodes in *Group 6*)

Table III
BINDING SITE RESIDUES OF CDK2 INTERACTING WITH THE 2 MOST POTENT SULFONAMIDE ANALOGUE INHIBITORS

Residue	Atom	GReMLIN
ASP145	CB	✓
	CG	✓
	OD1	•
LYS33	CB	✓
	CD	✓
	CE	✓
	CG	✓
	NZ	•
ASP86	N	✓
	CB	✓
	OD1	•
	OD2	•
LYS89	CB	✓
	CE	✓
	NZ	•
GLN85	-	×
HIS84	O	✓
LEU83	N	✓
	O	✓
PHE82	CE2	•
	CZ	•
GLU81	O	✓
PHE80	CB	✓
	CG	•
	CD2	•
	CE2	•
	CZ	•

Table IV
ACTIVE SITE RESIDUES OF RICIN CHAIN A INTERACTING WITH A CYCLIC TRANSITION STATE ANALOGUE INHIBITOR

Residue	Atom	GReMLIN
GLY121	O	•
ARG180	NH1	•
	NH2	•
VAL81	N	•
	O	•
ASP96	OD1	•
	OD2	•
ASP100	OD2	•
ASP75	OD2	•
ASN78	ND2	•
TYR80	CD1	✓
	CD2	✓
	CE1	✓
	CE2	✓
	CG	✓
	CZ	✓
TYR123	N	•
	CD2	✓
	CE2	✓
	CG	✓
GLU208*	-	×
GLU177	OE2	✓
ARG134*	-	×

✓ Residues/atoms found in patterns; • Found but not in patterns; × Not found.

shows atoms CD1, CG, CE1, CZ, CE2, all of them aromatic/hydrophobic from TYR80. In the pattern of 5 nodes from *Group 9*, we found CD2, CE2, and CG from TYR123. In the pattern of *Group 5* (2 aromatic/hydrophobic protein nodes connected to a hydrophobic ligand node) we found TYR80:CD2. It is important to point out that these residues can appear in other patterns from other groups (for example, TYR123 in *Group 7*). We are just spoting some examples from GReMLIN patterns involving relevant residues in ricin-ligand interaction according to the experimental results.

In *Group 5*, we see VAL81:N (donor, purple node). VAL81:O (acceptor, dark green) can be seen in *Group 1*. Also in *Group 1*, we see GLY121:O (acceptor, dark green node), ARG180 NH1 and NH2 (donor/positive, rose node), ASP96 OD1 and OD2 (acceptor/negative, red node), ASP100 OD1 and OD2 (acceptor/negative, red node), ASN78 ND2 (donor, purple node), GLU177:OE2 (acceptor/negative, red node), and TYR123:N (donor, purple node). All these mentioned atoms from *Group 1* do not appear in a frequent pattern. This group is large and heterogeneous, involving varied types of atoms, which makes difficult to find general frequent subgraphs with $support$ values greater than 0.6. One possible strategy to address this issue is considering another $support$ value for this kind of group, which would allow to obtain less frequent yet relevant patterns.

In [18], authors highlight TYR80 and TYR123 as relevant residues that establish π-staking (aromatic) interactions. More specifically, ARG180 at one end of the aromatic interaction provides cationic polarization and GLU177 serves to activate H2O nucleophiles. It is important to notice that GReMLIN was able to find all TYR80 and TYR123 atoms as frequent subgraphs (patterns), but TYR123:N, which appears in GReMLIN but not in a frequent subgraph. GLU208 and ARG134 do not directly interact with ligand [18]. These results are summarized in Table IV

IV. CONCLUSION AND FUTURE WORK

This work proposed GReMLIN, a strategy to search for conserved protein-ligand interactions based on frequent subgraph mining. The strategy models protein-ligand interfaces as bipartite graphs where nodes represent protein or ligand atoms and edges represent the interactions between them. Also, nodes are labeled with physicochemical properties of atoms they represent and edges are labeled according to their atoms' properties and a distance criterion. Next, a clustering analysis is conducted on the computed graphs and a frequent subgraph mining is performed on each cluster to find common subtructures conserved in protein-ligand interface all over a family. The results show that GReMLIN is able to find frequent substructures in protein-ligand interface (we illustrate this by discussing some patterns). Moreover, when compared to experimentally determined interactions, our *in silico* strategy was able to find many of relevant binding site residues for CDK and active site residues for Ricin.

As future work, we intend to enhance GReMLIN clustering strategy, as the results from FSM are highly dependent on the quality of clusters. For instance, *Group 1* from Ricin dataset has a diverse set of graphs regarding node types and graph topology, which makes difficult to find subgraphs with high frequency. Also, we want to consider structural water molecules to compute water mediated interactions. Finally, it is important to point out that GReMLIN found some patterns that do not involve binding/active site residues. We intend to further investigate whether such residues can be involved in allosteric regulation.

FUNDING

This work has been supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (Grant Number 477587/2013-5) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

REFERENCES

[1] M. F. Dunn, "Protein–ligand interactions: General description," *eLS*, 2007.

[2] D. E. Pires *et al.*, "acsm: noise-free graph-based signatures to large-scale receptor-based ligand prediction," *Bioinformatics*, 2013.

[3] H. Berman *et al.*, "The protein data bank," *Nucleic acids research*, vol. 28, no. 1, p. 235, 2000.

[4] S. Kim *et al.*, "Pubchem substance and compound databases," *Nucleic acids research*, p. gkv951, 2015.

[5] T. Sterling and J. J. Irwin, "Zinc 15-ligand discovery for everyone," *Journal of chemical information and modeling*, 2015.

[6] D. S. Wishart *et al.*, "Drugbank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D668–D672, 2006.

[7] R. A. Laskowski *et al.*, "Ligplot+: multiple ligand–protein interaction diagrams for drug discovery," *Journal of chemical inf. and mod.*, vol. 51, no. 10, pp. 2778–2786, 2011.

[8] H. Ding *et al.*, "Similarity-based machine learning methods for predicting drug–target interactions: a brief review," *Briefings in Bioinformatics*, p. bbt056, 2013.

[9] H. Yabuuchi *et al.*, "Analysis of multiple compound–protein interactions reveals novel bioactive molecules," *Molecular systems biology*, vol. 7, no. 1, p. 472, 2011.

[10] T. van Laarhoven, S. B. Nabuurs, and E. Marchiori, "Gaussian interaction profile kernels for predicting drug–target interaction," *Bioinformatics*, vol. 27, no. 21, pp. 3036–3043, 2011.

[11] M. Gönen, "Predicting drug–target interactions from chemical and genomic kernels using bayesian matrix factorization," *Bioinformatics*, vol. 28, no. 18, pp. 2304–2310, 2012.

[12] Y. Yamanishi, M. Kotera, M. Kanehisa, and S. Goto, "Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework," *Bioinformatics*, vol. 26, no. 12, pp. i246–i254, 2010.

[13] M. Kuhn *et al.*, "Stitch 4: integration of protein–chemical interactions with user data," *Nucleic acids research*, p. gkt1207, 2013.

[14] L. N. Johnson, "Protein kinase inhibitors: contributions from structure to clinical compounds," *Quarterly reviews of biophysics*, vol. 42, no. 01, pp. 1–40, 2009.

[15] E. Schonbrunn *et al.*, "Development of highly potent and selective diaminothiazole inhibitors of cyclin-dependent kinases," *Journal of medicinal chemistry*, vol. 56, no. 10, pp. 3768–3782, 2013.

[16] H. Y. Chen, L. Y. Foo, and W. K. Loke, "Ricin and abrin: A comprehensive review of their toxicity, diagnosis, and treatment," *Toxinology*, 2015.

[17] M. C. Bassik *et al.*, "A systematic mammalian genetic interaction map reveals pathways underlying ricin susceptibility," *Cell*, vol. 152, no. 4, pp. 909–922, 2013.

[18] M.-C. Ho *et al.*, "Transition state analogues in structures of ricin and saporin ribosome-inactivating proteins," *Proceedings of the National Academy of Sciences*, vol. 106, no. 48, pp. 20 276–20 281, 2009.

[19] A. Okabe *et al.*, *Spatial tessellations: concepts and applications of Voronoi diagrams*. Wiley, 2009, vol. 501.

[20] V. o. Gonçalves-Almeida, "Hydropace: understanding and predicting cross-inhibition in serine proteases through hydrophobic patch centroids," *Bioinformatics*, vol. 28, no. 3, pp. 342–349, 2012.

[21] V. Sobolev *et al.*, "Automated analysis of interatomic contacts in proteins." *Bioinformatics*, vol. 15, no. 4, pp. 327–332, 1999.

[22] A. L. Mancini, *et al.*, "Sting contacts: a web-based application for identification and analysis of amino acid contacts within protein structure and across protein interfaces," *Bioinformatics*, vol. 20, no. 13, pp. 2145–2147, 2004.

[23] L. Eldén, "Numerical linear algebra in data mining," *Acta Numerica*, vol. 15, pp. 327–384, 2006.

[24] C. Bécavin *et al.*, "Improving the efficiency of multidimensional scaling in the analysis of high-dimensional data using singular value decomposition," *Bioinformatics*, vol. 27, no. 10, pp. 1413–1421, 2011.

[25] S. de Azevedo Silveira *et al.*, "Enzymap: Exploiting protein annotation for modeling and predicting ec number changes in uniprot/swiss-prot," *PloS one*, vol. 9, no. 2, p. e89162, 2014.

[26] S. Deerwester *et al.*, "Computer information retrieval using latent semantic structure," Jun. 13 1989, uS Patent 4,839,853.

[27] X. Yan and J. Han, "gspan: Graph-based substructure pattern mining," in *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*. IEEE, 2002, pp. 721–724.

[28] M. Koyutürk *et al.*, "An efficient algorithm for detecting frequent subgraphs in biological networks," *Bioinformatics*, vol. 20, no. suppl 1, pp. i200–i207, 2004.

[29] X. Yan and J. Han, "Closegraph: mining closed frequent graph patterns," in *Proceedings of the ninth ACM SIGKDD*. ACM, 2003, pp. 286–295.

[30] J. Huan *et al.*, "Spin: mining maximal frequent subgraphs from graph databases," in *Proceedings of the tenth ACM SIGKDD*. ACM, 2004, pp. 581–586.

[31] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[32] L. P. Cordella *et al.*, "A (sub) graph isomorphism algorithm for matching large graphs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 10, pp. 1367–1372, 2004.

Apêndice B

Artigo Submetido

Alexandre V. Fassio, Santana, Charles A., Fabio R. Cerqueira, João P. R. Romanelli, Carlos H. da Silveira, , Raquel C. de Melo-Minardi, and Sabrina de A. Silveira. "**visGReMLIN: An interactive strategy to visualize commun subgraphs in protein-ligand interaction applied to human CDK2 and Ricin.**"Journal of Bioinformatics and Computational Biology, 2017.

Journal of Bioinformatics and Computational Biology

visGReMLIN: AN INTERACTIVE STRATEGY TO VISUALIZE COMMON SUBGRAPHS IN PROTEIN-LIGAND INTERACTION APPLIED TO HUMAN CDK2 AND RICIN

ALEXANDRE V. FASSIO*, CHARLES A. SANTANA†, FABIO R. CERQUEIRA†,
JOÃO P. R. ROMANELLI‡, CARLOS H. DA SILVEIRA‡,
RAQUEL C. DE MELO-MINARDI*, SABRINA A. SILVEIRA†

* *Computer Science Department, Universidade Federal de Minas Gerais*
Belo Horizonte, Minas Gerais 31270-901, Brazil
{*alexandrefassio, raquelcm*}*@dcc.ufmg.br*

† *Computer Science Department, Universidade Federal de Viçosa,*
Viçosa, Minas Gerais 36570-900, Brazil
{*charles.santana, fabio.cerqueira, sabrina*}*@ufv.br*

‡ *Universidade Federal de Itajubá,*
Itabira, Minas Gerais 35903-087, Brazil
{*joaoromanelli, carlos.silveira*}*@unifei.edu.br*

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

Interactions between proteins and ligands play an important role in biological processes of living systems. For this reason, the development of computational methods to facilitate the understanding of the ligand-receptor recognition process is of fundamental importance, since this comprehension is a major step towards ligand prediction, target identification, lead discovery, among others. This article presents visGReMLIN, a visual interactive interface to explore protein-ligand interactions and their conserved substructures for a set of similar proteins. To illustrate the potential of our strategy, we used two test datasets, Ricin and human CDK2, which have their protein-ligand interface modeled as bipartite graphs, where an edge depicts an interaction between a protein node and a ligand node. Such graphs are the input to search for frequent subgraphs that are the conserved interaction patterns over the datasets. The input graphs and their patterns can be explored to find general trends and exceptions concerning types of atoms and interactions. A text search to help users to find residues/atoms of interest (for example, atoms from CDK2 hinge region and from Ricin A chain active site) is also provided. Additionally, visGReMLIN provides some visualizations of basic statistics on frequencies of atoms and interactions of specific types in the dataset. Finally, our strategy permits users to select an interaction pattern to highlight it in the context of 2D interface graphs and in a 3D molecule viewer.

Availability: http://dcc.ufmg.br/~alexandrefassio/gremlin/

Keywords: visualization; pattern; protein; ligand; interaction; graph; data mining

1. Introduction

Protein-ligand interactions, which refer to noncovalent bonding such as aromatic stacking, hydrogen bonding, hydrophobic forces and salt bridges, play a crucial role in molecular recognition. In accordance with Ref. 21, the conditions responsible for the binding and interaction of two or more molecules are a combination of conformational and physicochemical complementarity. Hence, understanding, characterizing, and using knowledge of protein-ligand interactions can lead to target protein identification, prediction of hit as well as lead compounds and, ultimately, the determination of drug candidates[34,30].

Usual methods for *in silico* prediction of interactions between proteins and small molecules are classified into *ligand-based* and *structure-based* approaches. *Ligand-based* approaches generate or compare a candidate ligand to the known active molecules to identify compounds with similar bioactivity, whereas *structure-based* approaches use information about target structure to sample candidate molecules in target binding site [8]. Recently proposed techniques based on machine learning reached considerably success by taking the perpective of chemogenomics, which integrates attributes of drug compounds, proteins, and the known compound-protein interactions into a unified mathematical framework [27].

A remarkable motivation behind chemogenomics is that some classes of molecules can bind similar proteins, suggesting that the knowledge of some ligands for a target can be helpful to determine ligands for similar targets [20]. According to Ref. 20, there are some chemogenomic approaches to predict interactions between proteins and ligands, termed in Ref. 36: (i) *ligand-based*, which pools together targets at the level of families or subfamilies and learns a model for ligands at the family level [2,24]; (ii) *target-based*, which uses ligand binding site similarity to group receptors and pools together known ligands for each group to infer shared ligands; and (iii) *target-ligand* , which tries to predict ligands for a given target by leveraging binding information for other targets in a single step, without first attempting to define a particular set of similar receptors.

The mentioned approaches aim at predicting ligands or targets in a computational manner, which implies that there are some types of conserved patterns among similar ligands or receptors. In Ref. 39, we proposed GReMLIN (**Gr**aph **M**ining strategy to infer protein-**L**igand **IN**teraction patterns), a strategy to search for conserved protein-ligand interactions in a set of related proteins, based on clustering and frequent subgraph mining, which is able to perceive structural arrangements relevant for the protein-ligand interaction. However, we realized that although our graph-based strategy is able to infer protein-ligand interaction patterns, it fails on informing such patterns for two reasons: (i) it is difficult to understand a protein-ligand complex interface modeled as a graph without a visual representation of this graph, especially when we are considering many proteins at once; and (ii) given the interaction patterns, which are common substructures found in the protein-ligand interface, domain specialists would be interested in visualizing such interactions in

the context of protein structures, i.e., in a 3D molecule representation.

In this paper, we propose visGReMLIN, a visual interactive interface to explore protein-ligand interactions and their common substructures computed by GReMLIN. Interactive visualizations can be particularly interesting to represent complex and high volume data as well as to support users on revealing tendencies and exceptions in those data. Here, interactions and their patterns are modeled as graphs at the atomic level, where a vertex is a protein or ligand atom and edges depict interactions between atoms. Both vertices and edges are labeled with physicochemical properties of the entities they represent. We present interactions for two datasets, Ricin and human CDK2, where the interfaces of protein-ligand complexes are modeled as graphs in which color is a pre-attentive atributte that encodes physicochemical properties of atoms and interactions. Thus, users can see at a glance general trends and exceptions concerning types of atoms and interactions. Furthermore, we provide a variety of filters to explore interactions and their patterns as well as a text search to help users to find residues/atoms in which they are particularly interested. Finally, visGReMLIN allows to select an interaction pattern and highlight it in the context of 2D interface graphs and in a 3D molecule viewer. visGReMLIN was implemented in Data-Driven Documents[a] (D3).

2. Problem modeling

In this section, we explain how we compute conserved substructures in protein-ligand interfaces using the GReMLIN strategy. Figure 1 provides a schematic view of the GReMLIN workflow. It is divided in three main steps: *Graph dataset generation*, *Clustering analysis* and *Pattern computation*.

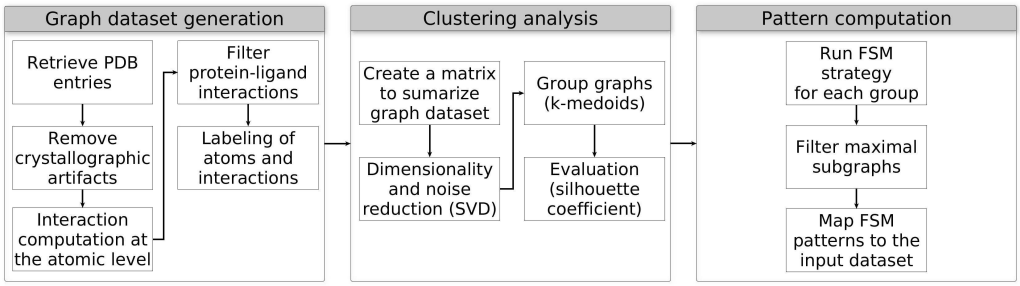


Fig. 1. GReMLIN workflow to calculate protein-ligand interaction patterns.The workflow is divided in three main steps: Graph dataset generation, Clustering analysis and Pattern computation.

Given a dataset (a collection of similar proteins), the first step is to represent the interface between proteins and their ligands as graphs in which atoms from proteins

[a]https://d3js.org/

and ligands are nodes and the interactions between atoms are edges. Ligands with 6 or less atoms were considered crystallographic artifacts and were thus removed as in Ref. 34. The interactions were computed through a Voronoy tessellation followed by a Delaunay triangulation[35,32] which is a cutoff-independent approach that avoids occlusion [7]. The interactions in the protein-ligand interface were filtered to the next steps.

Thereafter, nodes were labeled as *positively charged*, *negatively charged*, *aromatic*, *hydrophobic*, *donor*, or *acceptor* based on our previous work in Ref. 14, which, in turn, was derived from Ref. 44. Ligand nodes were labeled by the Pmapper software from Chemaxon (Pmapper 5.3.8, 2010, Chemaxon 2) at pH 7.0. According to predefined distance criteria[28,43] and the type of nodes, edges were labeled as *aromatic stacking*, *hydrogen bond*, *hydrophobic*, *repulsive*, and *salt bridge*. Nodes and edges can receive one or more labels . The character "/" is the separator of multiple labels associated to the same node or edge. So, for example, in Table 2, *aromatic/hydrophobic* means that there are nodes in group 1 that are labeled with 2 physichochemical properties at the same time, which are are *aromatic* and *hidrophobic*. Additionally, in group 1, there are nodes labeled with just one physicochemical characteristic, which is *hydrophobic*. The composite labels are separated by "**;**". Table 1 shows the distance criteria used.

Table 1. Distance criteria used to compute interactions (in Å).

Interaction type	Atom types	Min. distance	Max. distance
Aromatic stacking	2 aromatic atoms	1.5	3.5
Hydrogen bond	1 acceptor and 1 donor atom	2.0	3.0
Hydrophobic	2 hydrophobic atoms	2.0	3.8
Repulsive	2 atoms with the same charge	2.0	6.0
Salt bridges	2 atoms with opposite charge	2.0	6.0

At this point, the protein-ligand interfaces built from proteins of the input dataset are represented as bipartite graphs in which every edge connects a protein node to a ligand node. Next, a clustering analysis is performed on the dataset of graphs in order to search for frequent subgraphs in each group. To group the graph dataset, it is summarized by a matrix in which each row represents a graph and each column represents a pair of node types in the end of an edge. In other words, the columns depict all interaction types. Each entry in this matrix contains the amount of a specific interaction type for a certain graph. The dimensionality and noise reduction of the matrix is performed by using Singular Value Decomposition (SVD)[33]. The resulting matrix is grouped using the K-medoids[23] algorithm with all possible values for K (number of groups) and the best clustering result is chosen according to the silhouette coefficient[37]. It is known that the singular vectors in the SVD define a new orthogonal space so that their directions are extended according

to the maximum variations of the data. Singular values have the role of scaling the singular vectors to best represent these variations. In our case, attributes with greater variations tend to be associated with the first (greater) singular vectors, so that we can give them a "meaning" in function of the physicochemical properties of atoms and contacts. Consequently, if each singular vector represents a latent attribute, then it is also possible to divide it into subgroups according to the intensity of that attribute. Thus, two rounds of clustering were made: a first, to define groups according to the major singular vectors; a second, to define subgroups within each of them.

Once the dataset of graphs is segmented in groups, a Frequent Subgraph Mining (FSM) experiment is conducted using the algorithm gSpan[49] to extract frequent subgraphs that represent the common substructures conserved in the protein-ligand interfaces of each group. Next, we filter only the maximal subgraphs for the subsequent steps. A subgraph g is maximal frequent if it satisfies the conditions[18]: (i) g must be frequent, and (ii) there must be no frequent super graph of g. In accordance with Ref. 25, in biological networks, maximal frequent subgraphs are deemed to be the most interesting ones. The resulting maximal subgraphs, which we call patterns, are then mapped to the graph input dataset with the VF2[5] subgraph isomorphism algorithm. This mapping step is necessary because the FSM algorithm computes frequent subgraphs but does not show where these subgraphs are in the input dataset.

Algorithm 1 shows the functions that calculate the GReMLIN protein-ligand interaction patterns. To compute such patterns, the algorithm takes as input the following parameters: A set of proteins from Protein Data Bank [3] (PDB), which we named proteinSet, a list with physicochemical properties of each atom from each residue (atomClass), and the minimum number of atoms to consider a non-protein small molecule as a ligand (atomCutoff). In line 1, we define the prototype of the InterfaceAsGraphs function. In line 2, we remove cristalographic artifacts from the protein set. Contact computation between atoms from each PDB entry is performed in line 3 using Delaunay, resulting in a set of graphs which represents atomic contacts for the whole PDB entry. In line 4, the set of contactGraphs is filtered to keep only contacts involving a protein and a ligand atom. Furthermore, nodes and edges of these graphs are labeled with their physicochemical properties. In line 7, we define the prototype of the GReMLINpatterns function. The function InterfaceAsGraphs is called in line 8 to model the protein set as graphs that represent contacts in the protein-ligand interface. In line 9, we perform a clustering analysis in the graph dataset. We iterate over each group of graphs in line 10. The FSM experiment is conducted over each group of graphs in line 11. In line 12, we filter only the maximal subgraphs for the subsequent step. We map the patterns (common maximal subgraphs), resultant from the FSM experiment, to the graph input dataset (groupsOfGraphs) in line 13.

This modeling of interactions in the protein-ligand interface as graphs as well as the clustering analysis and frequent subgraph mining that computes conserved

interaction patterns are part of our work in Ref. 39. Here, we make a brief review of the necessary concepts to understand the visualization strategy we propose in this paper to explore interactions in the protein-ligand complex interface and their patterns.

Algorithm 1 GReMLIN patterns calculation

```
1: function InterfaceAsGraphs(proteinSet, atomClass, atomCutoff)
2:     proteinSet ←removeCristalographicArtifacts(proteinSet, atomCutoff)
3:     contactGraphs ←calculateDelaunayAtomicContacts(proteinSet)
4:     contactGraphs ←filterLabelPLInteractions(contactGraphs, atomClass)
5:     return contactGraphs
6: end function

7: function GReMLINpatterns(proteinSet, atomClass, atomCutoff)
8:     graphs ←InterfaceAsGraphs(proteinSet, atomClass, atomCutoff)
9:     groupsOfGraphs ← groupGraphsKMedoids(graphs)
10:    for all group ∈ groupsOfGraphs do
11:        patternSet[group] ← runFSM(group)
12:        patternSet[group] ← filterMaximalSubgraphs(group)
13:        patternSet[group] ← mapFSM(patternSet[group], groupsOfGraphs)
14:    end for
15:    return patternSet
16: end function
```

3. Dataset

Our datasets were downloaded from PDB and comprise two sets of similar proteins in which we are interested in extracting protein-ligand interaction patterns. Further details about datasets can be found in visGReMLIN web site and in Ref. 43.

- CDK2: This dataset, which is based on the work in Ref. 40, comprises a specific protein for which several inhibitors are known. This same protein was crystallized with a variety of ligands, and the 73 experimental structures, with identical sequences, are available in PDB. Also, in Ref. 40, authors described the development of highly potent and selective diaminothiazole inhibitors of CDK2 based on a single hit compound with weak inhibitory activity. We extracted from that work binding site residues and atoms experimentally determined that are relevant for CDK2 interaction with ligands.
- Ricin: It is composed of 29 experimental structures from PDB, which have at least one ligand and 50% or more identity with ricin A chain (2AAI:A in PDB). We consider this dataset a more realistic one, as the sequences

are not exactly the same, which is common, for instance, in a protein family. In Ref. 17, authors co-crystalized ricin chain A with a transition state analogue inhibitor that mimics the sarcin-ricin recognition loop of the eukaryote 28S rRNA. We extracted from such work active site residues and atoms experimentally determined that are relevant for the interaction of ricin chain A with subunit 28S rRNA.

4. Related work

In this section, we review some representative strategies proposed for visual representations of protein-ligand interactions.

Some strategies focus on representing protein-ligand complexes as 2D diagram in which the ligand and the interacting residues are depicted in a static way. LIGPLOT[47] is one of the first tools to take advantage of such strategy and today its results are used by other protein-ligand interaction tools (PLIC[1], PDBSum[9], PLI[13], PDBe[46]). PoseView is presented in Ref. 45 and also focus on the 2D representation of the interaction network between the complex partners. However, LIGPLOT and PoseView do not allow to compare protein-ligand interactions from different complexes. In Ref. 4, schematic diagrams for one or more complexes can be ploted in a static way, which allows comparison of interactions based on superposition of 3D structures to identify equivalences between ligands and interacting residues. Similarly, LigPlot^{+} [26] is a tool which generates interactive 2D protein-ligand interaction diagrams and allows to superimpose these diagrams for a few complexes. Furthermore, it highlights conserved interactions for protein residues that are in equivalent 3D positions when the two structures are superposed and show the 3D visualization in molecular viewers. Although sc-PDB[10] is a curated database of binding sites, it also presents 2D diagrams generated by PoseView. Also, LigDig[12] is a web server for querying protein-ligand interactions which also provides 2D representation of these interactions though it is not its main focus

Other tools like CREDO[41], GIANT[22], HET-PDB Navi[48], PLIP[38], BINANA[11], STING[29], Relibase [16] and MAESTRO[42] represent protein-ligand complexes as 3D representations using molecular viewers such as JSMol[15], GLmol[31] or VMD[19].

Here, we are interested in delivering a scalable interactive strategy to visualize protein-ligand interactions and their patterns accross a dataset of similar proteins, allowing users to explore, filter, and highlight conserved interaction patterns in schematic 2D representations or in a 3D molecule viewer. Also, we provide some visualizations of the frequencies of atoms and interactions of specific types in the dataset.

5. visGReMLIN

The main objetive of visGreMLIN is to present a panoramic view of protein-ligand interaction patterns for a set of related proteins. Also, we want to permit users to explore these patterns in detail by size, physicochemical type of atoms, and

interactions. Furthermore, the user can pose and solve general questions about conserved protein-ligand interactions.

Concerning this objective, we present the subgraphs that represent interaction patterns and their details accross four main sections presented below: *Dataset details*, *Graph patterns table*, *Graph patterns view* and *Graphical analysis*.

5.1. *Dataset details*

General information for a specific dataset of related proteins is displayed as a table (Fig. 2). In the column *Group*, we show the group from the clustering analysis. The graph identifiers coupled with their PDB entries and chains belonging to the respective group are provided in *PDB ids* column. By clicking on the *Points* icon, 2D representations for all ligands from a group are exhibited and, by clicking on the *Hexagon* icon, graph representations for all protein-ligand interfaces from a group are provided. Users can click on PDB entries to be directed to the PDB web site. We also provide a text search on this table.

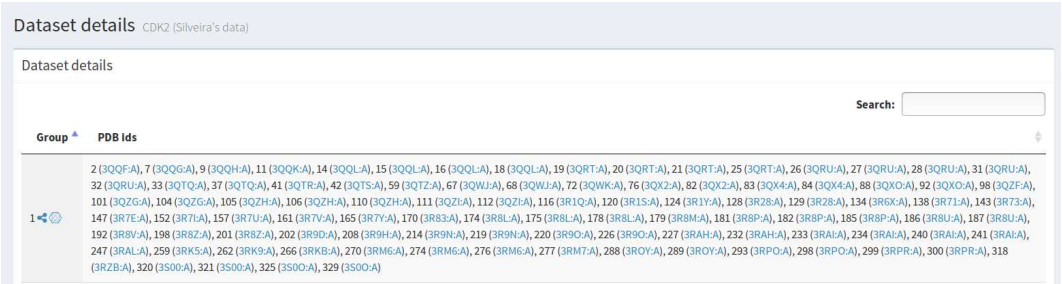


Fig. 2. Dataset details table.

5.2. *Graph patterns table*

A summary information regarding the *frequent subgraph minig experiment* is provided in this section. There are two kinds of tables, which we named *Grouping columns* and *Simple table*. In *Grouping columns*, Fig. 3, the column *Pattern size* displays the number of nodes for each type of pattern (subgraph). The column *Ocurrences* shows the number of types of patterns (subgraphs) with a specific size (number of nodes). This table is also segmented by *Group* from the clustering analysis and *Support* value and it is an interactive table, in which by clicking on subheaders *Group* and *Support* data are ordered in ascending or descending order. Additionally, we deliver some filters to explore this table by group, by minimum pattern size, and by minimum occurrences.

In *Simple table*, Fig. 4, we use a heatmap representation in which color is a pre-attentive attribute that encodes the frequency of subgraphs in a shade of blue.

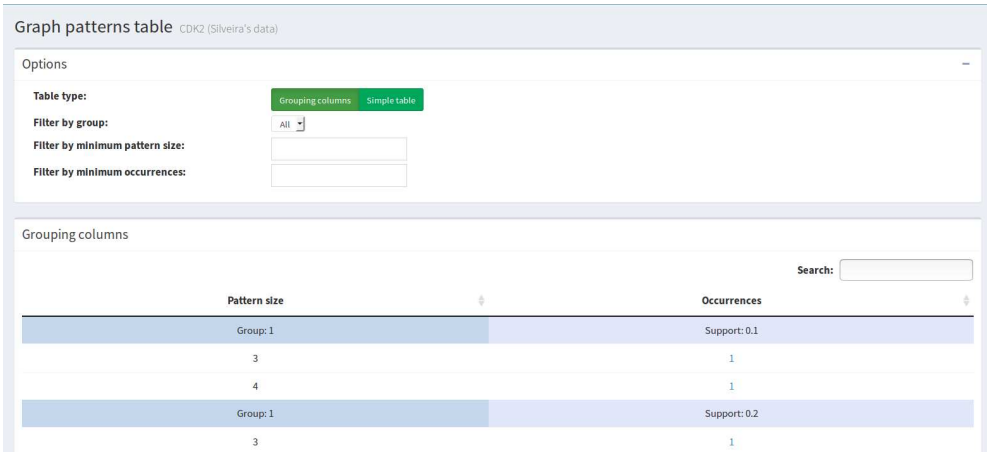


Fig. 3. Grouping columns.

Subgraphs are segmented by *size* (number of nodes), which is displayed in the first column, as well as by *group* (from the clustering analysis) and *support value* (from frequent subgraph mining). For instance, in the top left blue position from CDK2 *Simple table*, we see the number 1, which means that, for support value 0.1 (top header line), in group 2 (second header line), there is 1 type of subgraph (pattern) whose size is 2 (label in column 1). This table helps choosing an appropriate support value for mining protein-ligand interaction patterns. As the support increases, we lose some patterns. Therefore, by inspecting this table, we can choose the support value that helps us keeping interesting patterns (many nodes).

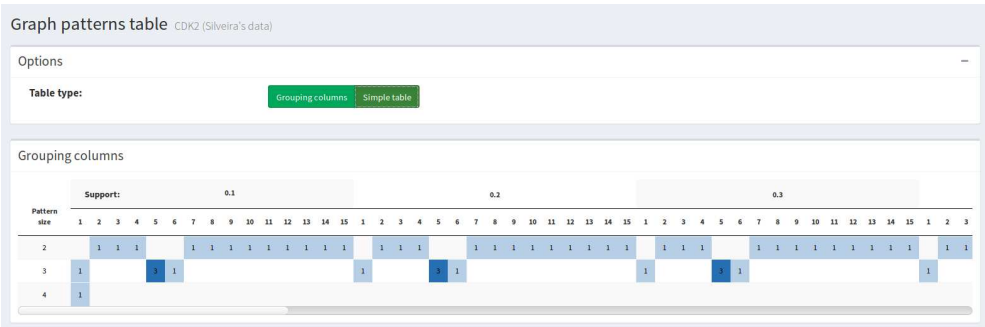


Fig. 4. Simple table.

5.3. *Graph patterns view*

This is the most important section of our visualization tool, as it depicts the computed patterns in the protein-ligand interface. It is organized in three subsections, which allow users to perform analytical interaction and navigation all over the frequent subgraphs. An image from this section of visGReMLIN is provided in Fig. 5.

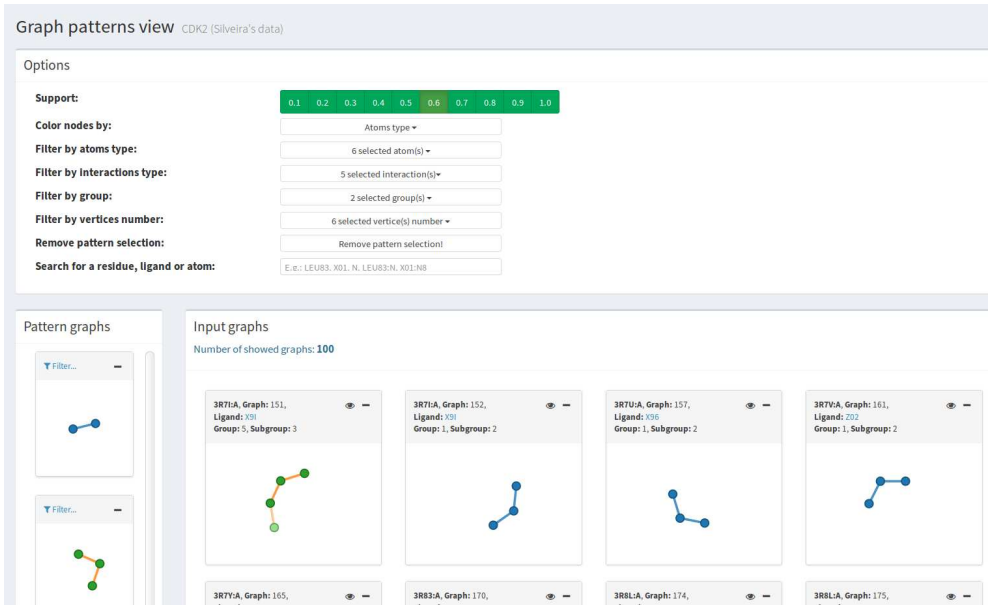


Fig. 5. Graph patterns view.

The subsection *Options* permits to interact with the patterns through filters. The common workflow is choosing a support value (based on *Graph patterns table*) and explore subgraphs using the filters below:

Color nodes by: Nodes are coloured according to atom type (one color for each type) or molecule (one color for atoms from proteins and another color for atoms from ligands).

Filter by atom type: Atoms of the selected type are highlighted. Possible types are acceptor, aromatic, donor, hydrophobic, negative and positive.

Filter by interaction type: Interactions of the selected type are highlighted. Possible types are aromatic stacking, hydrogen bond, hydrophobic, repulsive and salt bridge.

Filter by group: Only graphs from the selected groups from clustering analysis are displayed.

Filter by vertex number: Only graphs with the selected number of vertices are

displayed.

Remove pattern selection: If a pattern was selected in the *Pattern graphs* section, it removes the selection, which displays all subgraphs according to the filters from subsection *Options*.

Search for a residue, ligand, or atom: Vertices from graphs that contain residue/ligand/atom in the text search are highlighted.

In the *Pattern graphs* subsection, the users can navigate through patterns, which are the frequent subgraphs for a dataset of graphs representing a set of related proteins. By clicking on a pattern, only subgraphs that contain such pattern are displayed on the subsection *Input graphs*. By passing the mouse over nodes and edges, we see the types of atoms and interactions on demand. Only patterns from groups selected in *Filter by group* are shown. This mapping from frequent subgraphs to the original input graphs is performed by using the VF2[5] algorithm implemented in the igraph[6] package for the C++ programing language.

The *Input graphs* subsection displays the graphs that represent interactions on the protein-ligand interface for a set of related proteins according to filters from the section *Options* and pattern selected in *Pattern graphs*. For each graph, we show its PDB id and chain, ligand id, graph id, and group. By passing the mouse over the graph, we give the details below on demand for nodes and edges:

Protein atoms: Name and number of residue to which the atom belongs, atom name, chain, physicochemical type of atom. An example is provided in Fig. 6(a).

Ligand atoms: Ligand name and number inside PDB file, atom name, chain, physicochemical type of atom. For instance, see Fig. 6(b).

Interactions: Information about connected atoms (residue or ligand name, number, and atom), physicochemical type of interaction and distance between connected atoms in angstrons (Å). Fig. 6(c) shows an example of information provided for interactions.

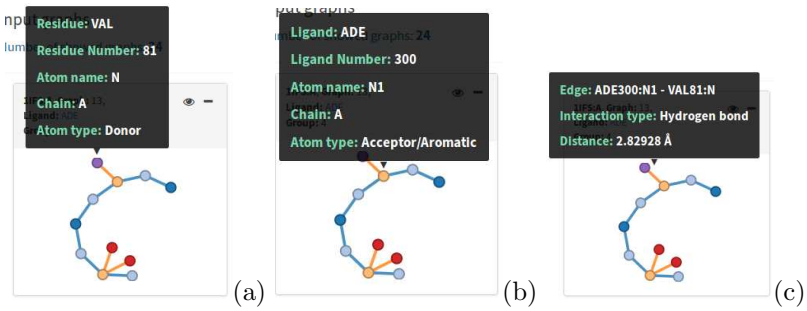
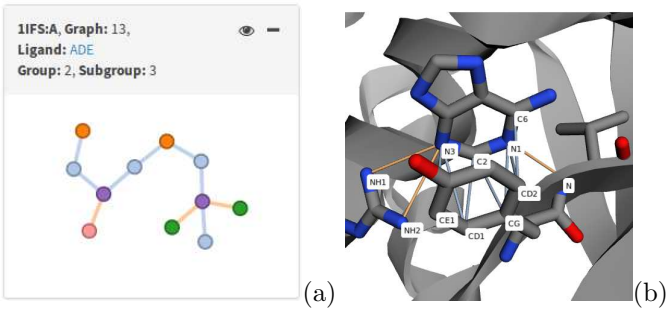


Fig. 6. Data displayed in the *Input graphs* section by passing the mouse over graphs for protein node (a), ligand node (b), and interaction (c), respectively.

In addition to a graph schematic 2D visualization, to support users on understanding the patterns in the context of protein structure, we provide a 3D representation of the protein-ligand interaction graphs in a molecule viewer by clicking on the *eye* icon. An example of this functionality can be seen in Fig. 7, which provides a graph schematic visualization for 1IFS:A interacting with ligand ADE in the Ricin dataset (Fig. 7(a)) and a molecular structure 3D representation for this graph (same PDB entry and interactions) (Fig. 7(b)).



(a) (b)

Fig. 7. Graph schematic 2D visualization (a) and the corresponding 3D protein structure representation in the molecule viewer (b).

We also provide a general 2D visualization for ligands by clicking on the ligand name in any graph from the subsection *Input graphs*. Only ligands from graphs displayed in subsection *Input graphs* are shown in the set of ligands. This visualization allows users to compare and contrast ligands, revealing global trends among them for specific groups. An example of this functionality is provided in Fig. 8.

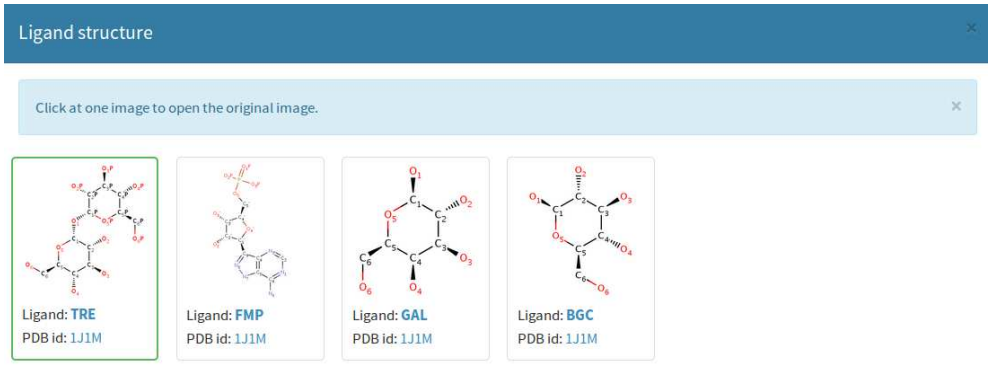


Fig. 8. Ligand structure in 2D representation for group 19 from the Ricin dataset.

5.4. *Graphical analysis*

In addition to the interactive visual representations for conserved interactions between protein and ligands in a set of related proteins, visGReMLIN provides an interactive interface to show frequencies of atoms and interactions of specific type. The common workflow in this section is choosing *Atoms type* or *Interactions type* tab and then selecting a physicochemical type of atoms or interactions. Then, a histogram will be displayed. Also, histogram bars can be grouped by *Support* value used in frequent subgraph mining or by *Group* (cluster) from the clustering analysis. In Fig. 9, we have an example of histogram for the CDK2 dataset, that shows the distribution of number of atoms (nodes from graphs) labeled as hydrophobic with bars grouped by cluster (Group). In this case, each bar represents the number of hydrophobic atoms for a specific support value. By passing the mouse over the bars, we provide the support and frequency values. It can be noticed that we have the majority of hydrophobic atoms in group 1 for all support values, while groups 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 15 do not have hydrophobic atoms. The same data is displayed in Fig. 10, with bars grouped by support value.

Fig. 9. Frequency of hydrophobic atoms for the CDK2 dataset grouped by cluster in the *Graphical analysis* section.

Fig. 10. Frequency of hydrophobic atoms for the CDK2 dataset grouped by support in the *Graphical analysis* section.

6. Results and discussion

6.1. *Pattern analysis*

6.1.1. *CDK2*

Clustering analysis for the CDK2 dataset resulted in 15 groups. Table 2 provides the number of graphs and the types of atoms present in each group. Users can explore each group by selecting the desired group in *Filter by group*. Visually inspecting groups for the CDK2 dataset, in general, we perceive that each one has a color signature. In other words, due to the use of color pre-attentive attribute in nodes and edges, users can see at a glance that the atom and interaction types are similar inside each group. On the other hand, visually comparing different groups, we notice that each one has a different color signature, which means that different groups involve different types of atoms and interactions. This indicates that our groups meet an important requirement of clustering analysis: High intra-cluster similarity and low inter-cluster similarity.

To illustrate the CDK2 results, we will discuss groups 1 and 2 as they are the most numerous groups. These groups together correspond to 59% of graphs from the CDK2 dataset. Group 1 is very homogeneous as the majority of nodes are hydrophobic. There is just one more type of node, which is aromatic/hydrophobic. As expected, this group has only one type of pattern, involving two hydrophobic nodes connected by a hydrophobic edge. From the 96 graphs on group 1, 96 have

Table 2. Number of graphs and types of nodes for each group of graphs from the CDK2 dataset.

Groups	# of graphs	Node types
1	96	hydrophobic; aromatic/hydrophobic
2	106	acceptor; donor; acceptor/donor; acceptor/negative;
3	75	acceptor; donor; acceptor/donor
4	18	donor; acceptor/donor; acceptor/negative
5	4	acceptor/negative; donor/positive
6	5	hydrophobic; aromatic/hydrophobic
7	5	acceptor/aromatic; donor/positive
8	10	donor; acceptor/aromatic
9	6	acceptor/donor; donor/positive
10	2	aromatic; aromatic/hydrophobic;
11	8	acceptor; donor/positive
12	1	donor; acceptor/donor
13	1	acceptor/negative; donor/positive
14	3	aromatic/hydrophobic
15	1	donor/positive; donor/positive

the mentioned pattern. Figure 11 shows some graphs present in group 1 and the only pattern of this group at the top-left corner.

Group 2 has mainly acceptor and acceptor/donor nodes. There are a few donor nodes and just one acceptor/negative node. The only pattern of this group is composed by one acceptor and one acceptor/donor node connected by a hydrogen bond. From the 106 graphs on this group, all of them have the mentioned pattern, which is shown in the top-left corner of Figure 12. This Figure also shows some graphs in group 2.

It is important to point out that by clicking on the patterns in visGReMLIN, these patterns are highlighted in each graph (that presents such pattern) on the right-hand side in a darker shade of the original color, while nodes and edges that are not part of the pattern appear in a lighter shade of the original color. In Figures 11 and Figure 12, we highlight this feature of our tool with blue rectangles.

6.1.2. *Ricin*

For the Ricin dataset, the clustering analysis resulted in 21 groups. The number of graphs and the types of atoms for each group are provided in Table 3. By visually inspecting Ricin dataset groups, we note a color signature for each group, which means that colors of nodes and edges are similar inside a group. Conversely, by comparing different groups, we note that they have different color signatures. Similarly to the CDK dataset, this indicates that groups have high intra-cluster similarity and low inter-cluster similarity.

However, it is important to point out that Ricin is a dataset smaller than CDK2 and, even so, it resulted in a higher number of groups in the clustering analysis, which indicates that, although CDK2 contains more graphs, they are more homo-

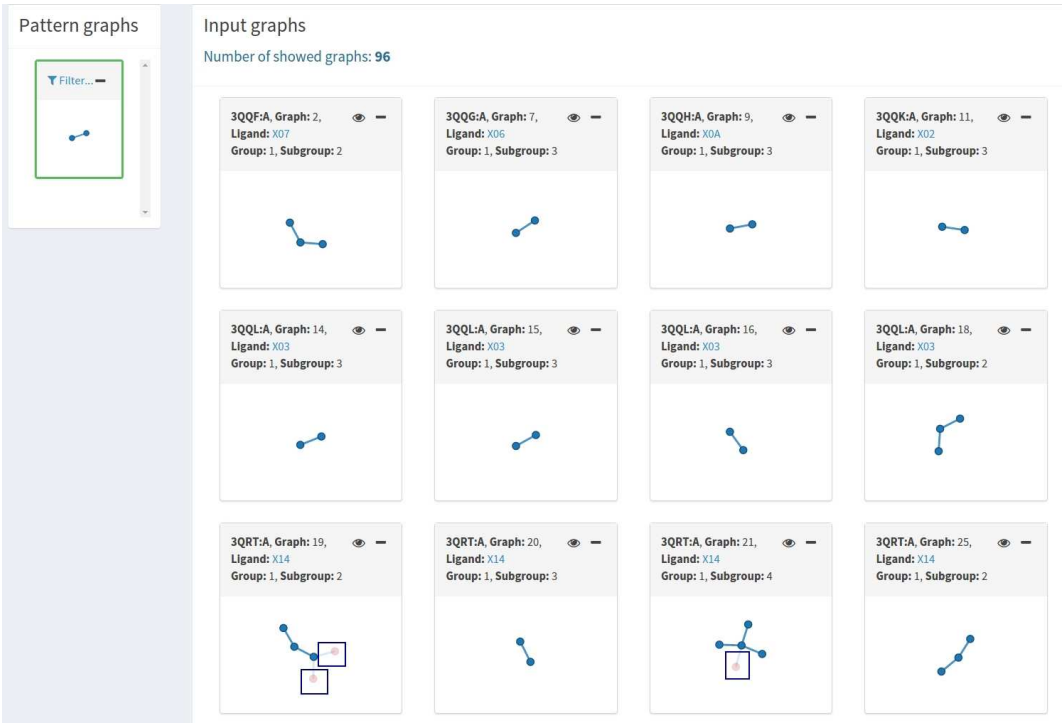


Fig. 11. Group 1 from the CDK2 dataset. The majority of atoms and interactions are hydrophobic and the group has 1 pattern (top-left corner) which occurs in all graphs of such group. We manually inserted the rectangles to show some nodes and edges that are not part of the highlighted pattern.

geneous than those from Ricin. The Ricin dataset contains 29 PDB entries that resulted in 197 protein-ligand interface graphs, while the CDK2 dataset is composed of 73 entries that resulted in 341 graphs. We consider CDK2 as a controlled scenario, as all structures have identical sequences with different ligands. Ricin is a more difficult and realistic scenario, as it involves a set of structures with different sequences, which happens, for instance, in protein families. Moreover, as Ricin is a more diverse set compared to CDK2, it is reasonable to consider that the clustering analysis of Ricin is more chalenging for data mining algorithms.

We discuss groups 1 and 2 from the Ricin dataset to illustrate our results, as these groups are the most numerous, with 28 and 30 graphs, respectively, which means about 30% of the dataset. Group 1 is considerably homogeneous as the majority of its nodes are hydrophobic or aromatic/hydrophobic. There are a few aromatic, donor and acceptor/aromatic nodes. This group has just one type of pattern composed of one hydrophobic node connected by hydrophobic interactions to 2 aromatic/hydrophobic nodes. Out of 28 graphs of this group, 20 have the mentioned pattern. Figure 13 shows the only pattern of group 1 and some of its

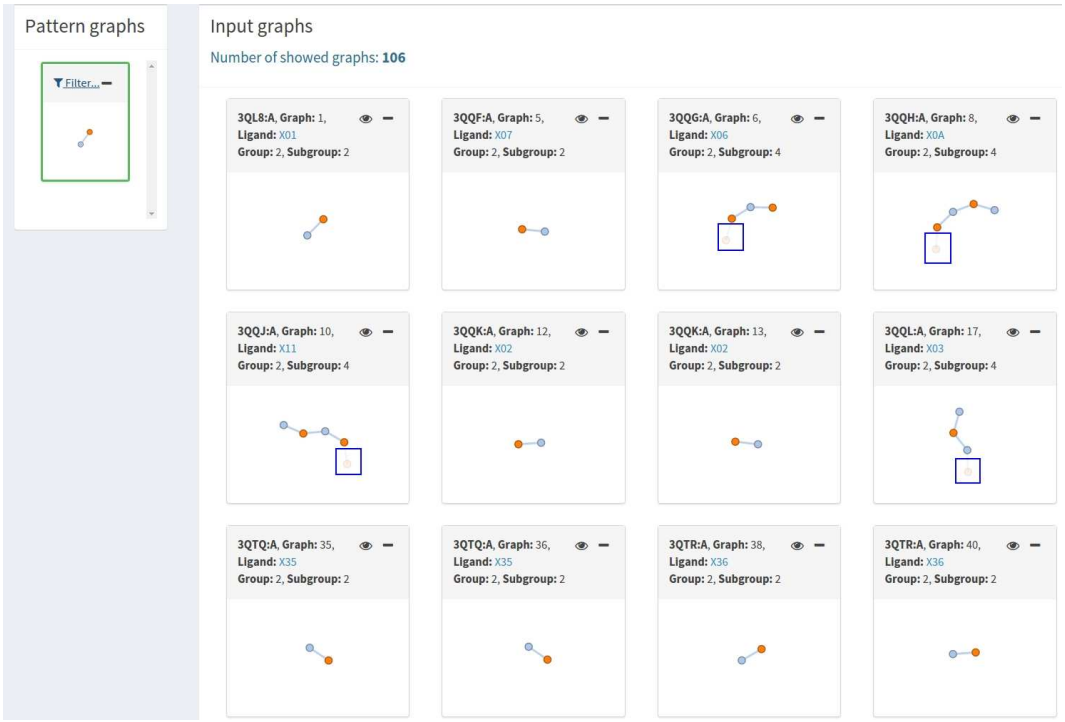


Fig. 12. Group 2 from the CDK2 dataset. The majority of atoms are acceptor and acceptor/donor and, consequently, interactions are mainly hydrogen bonds. We manually inserted the rectangles to show some nodes and edges that are not part of the highlighted pattern.

graphs.

There are mainly aromatic and aromatic/hydrophobic nodes in group 2, but a few graphs contain other types of nodes (acceptor, donor, hydrophobic, acceptor/aromatic, acceptor/donor, aromatic/donor, acceptor/aromatic/donor, donor/positive). This group has one pattern that is composed by one aromatic node and one aromatic/hydrophobic node connected by an aromatic interaction. Out of 30 graphs of group 2, all of them have the mentioned pattern. Figure 14 shows some graphs of group 2 and the only pattern of this group at the top-left corner.

Rectangles in Figures 13 and 14 show nodes in a lighter shade of their original color, indicating that such nodes are not part of the selected pattern at the top-left corner. visGReMLIN highlights patterns in the dataset of graphs by keeping the original color of nodes and edges that are part of the pattern.

6.2. *visGReMLIN patterns compared to experimental patterns*

We compare patterns computed through the GReMLIN strategy with relevant patterns experimentally determined for CDK2 and Ricin according to Ref. 40 and Ref.

Table 3. Number of graphs and types of nodes for each group of graphs from the Ricin dataset.

Groups	# of graphs	Node types
1	28	aromatic; donor; hydrophobic; acceptor/aromatic; aromatic/hydrophobic
2	30	acceptor; aromatic; donor; hydrophobic; acceptor/aromatic; acceptor/donor; aromatic/donor; aromatic/hydrophobic; acceptor/aromatic/donor; donor/positive
3	15	aromatic; donor; acceptor/negative; acceptor/aromatic; aromatic/hydrophobic; donor/positive
4	22	acceptor; aromatic; donor; acceptor/donor; aromatic/donor; aromatic/hydrophobic; donor/positive; acceptor/aromatic/donor
5	14	acceptor; donor; acceptor/donor; acceptor/negative; donor/positive
6	4	acceptor/aromatic; aromatic/hydrophobic; donor/positive; donor
7	12	acceptor; donor/positive
8	4	donor; acceptor/negative
9	8	donor; acceptor/aromatic
10	3	acceptor/negative
11	11	acceptor/aromatic; donor/positive
12	1	acceptor/donor; donor/positive
13	2	aromatic/donor; acceptor
14	2	acceptor; donor; aromatic/hydrophobic; acceptor/aromatic/donor
15	4	acceptor; acceptor/donor
16	12	acceptor; donor
17	12	hydrophobic
18	3	acceptor; donor; acceptor/donor
19	6	acceptor/donor; acceptor/negative
20	2	acceptor/donor
21	2	acceptor/donor; acceptor/negative; aromatic/donor/positive

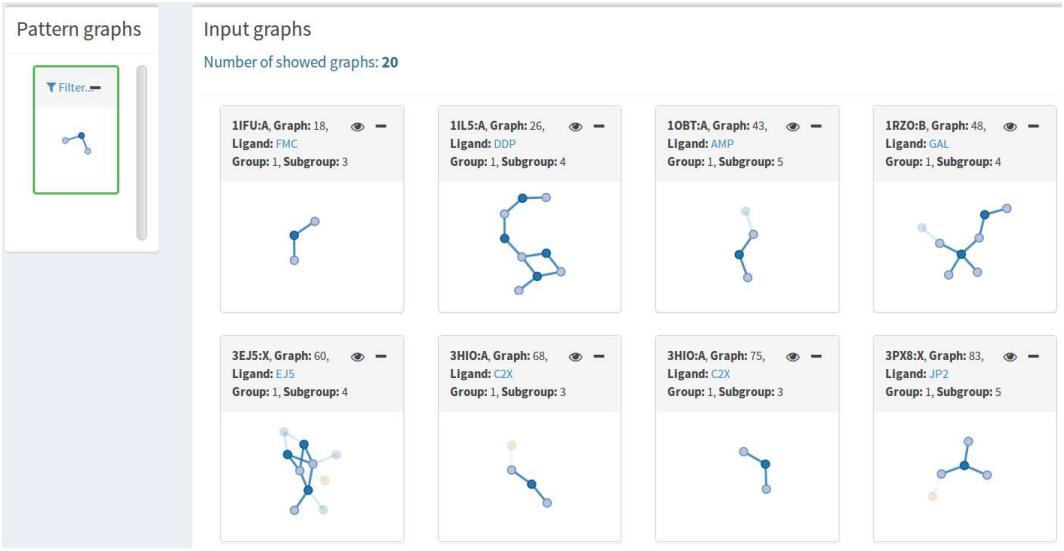


Fig. 13. Group 1 from the Ricin dataset. The majority of its nodes are hydrophobic or aromatic/hydrophobic and group 1 has 1 type of pattern (top-left corner) which occurs in 20 out of 28 graphs.

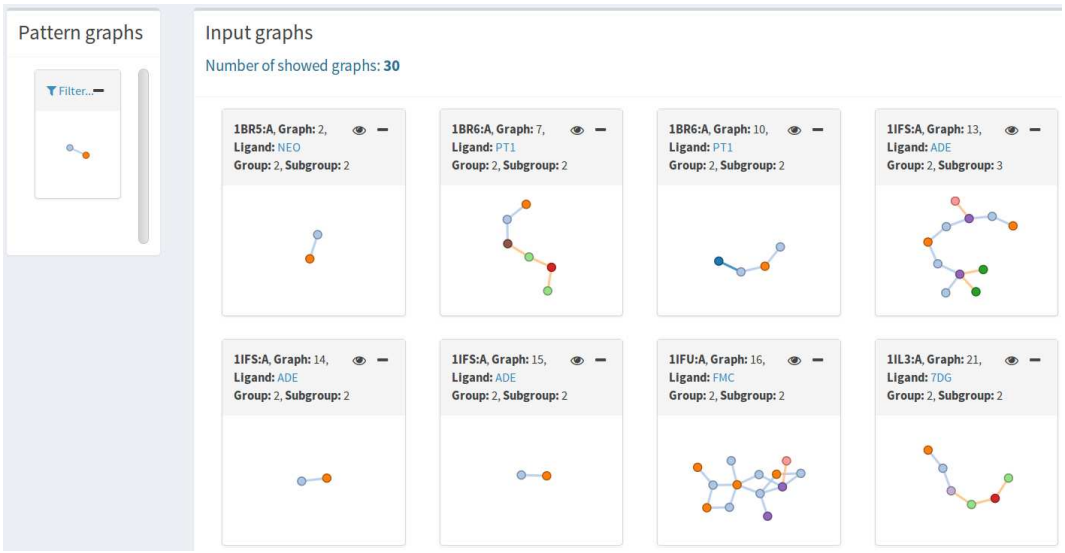


Fig. 14. Group 2 from the Ricin dataset. The majority of atoms are aromatic and aromatic/hidrophobic and, consequently, interactions are mainly aromatic. This group has 1 type of pattern, showed in the top-left corner, which occurs in all input graphs.

17, respectively, to verify whether our strategy is able to find the experimentally determined patterns. This is a qualitative analysis, as the residues determined as relevant in protein-ligand interactions for both studies do not represent interactions between a protein and all its possible ligands in the datasets used. However, we believe it is an interesting comparison, as these studies experimetally determined protein-ligand interactions stablished by Ricin and CDK2 with ligands that are very important for both proteins. It is important to point out that we show some examples of experimental patterns found by GReMLIN and that such patterns can be found in other groups and patterns of our strategy beyond those discussed here. Each graph has a unique identifier. When an atom of interest appears in only one graph of a group, we mention its specific identifier. We use pattern and common substructure as synonyms.

6.2.1. *CDK2*

We consider the set of binding site residues of CDK2 that interact with the 2 most potent sulfonamide analogue inhibitors developed in Ref. 40 as the experimentally determined patterns for the CDK2 dataset. Table 4 details these residues and which of them are detected using GReMLIN. Next, we discuss some representative patterns.

All atoms from ASP145 are present in GReMLIN patterns. Atoms CG and CB are represented in the only pattern of group1. Atom OD1 appears in the only

pattern of group 4. To check these atoms in the visualization tool, the user needs to select the desired group (1 or 4, in this case) in *Filter by group*, click in the pattern on the left-hand side in *Pattern graphs*, and write the desired residue and atom (ASP145:OD1, ASP145:OD1 or ASP145:OD1) in the text box in *Search for a residue, ligand or atom*. This process highlights nodes representing atoms from ASP145 by increasing their size.

Atoms CB, CD, CE and CG from LYS33 are depicted in the only pattern of group 1. Atom NZ from LYS33 is represented in the only pattern of group 7. In Figure 15, we provide an example where atoms CD, CE, CG from LYS 33 are in the graph representing the interaction between ligand X42 and protein 3QTZ:A. Figure 16 shows LYS33:NZ in group 7.

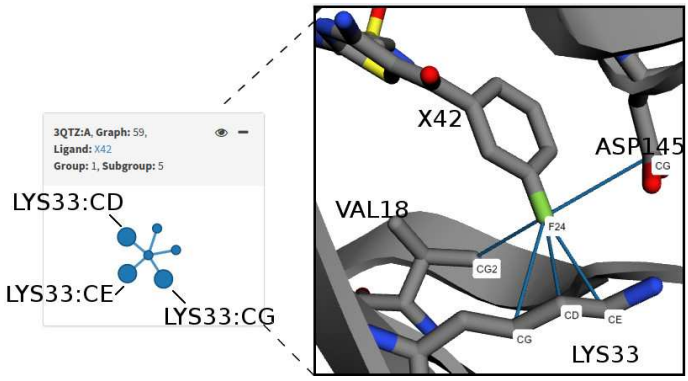


Fig. 15. Atoms CD, CE and CG from LYS 33 in 2D graph and in a molecule viewer, graph 59, in group 1, for the CDK2 dataset.

ASP86:N (donor) is in the only pattern of group 3. ASP86:CB (hydrophobic) appears in the only pattern of group 1 (in graph 37). ASP86:OD1 (acceptor/negative) appears in two patterns from group 5 and ASP86:OD2 (acceptor/negative) is in the only pattern of group 4.

Atoms CB and CE (both hydrophobic) from LYS89 are present in the only pattern of group 1, in graphs 241 and 259, respectively. LYS89:NZ (donor/positive) appears in 2 patterns of group 5.

HIS84:O (acceptor) is in the pattern of group 2.

LEU83:N (donor) is in the common substructure of group 3. LEU83:O (acceptor) is in the pattern of group 2.

Atoms CE2 and CZ from PHE82 (both aromatic/hydrophobic) are in group 1, but they are not in a common substructure.

GLU81:0 (acceptor) is in the pattern of group 2.

PHE80:CB (hydrophobic) is in the only pattern of group 1 and PHE80:CD2 (aromatic/hydrophobic) is in the pattern of group 14. Atoms CG, CE2 and CZ

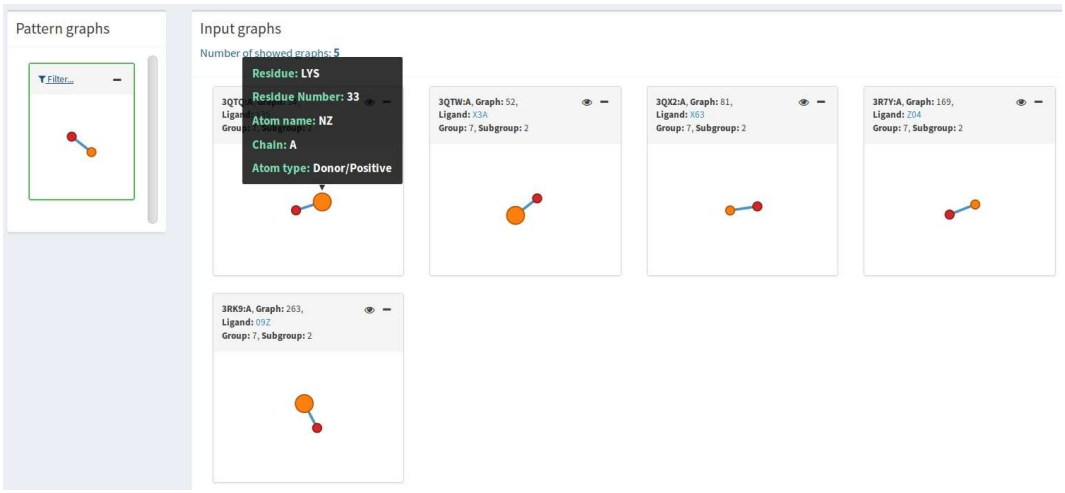


Fig. 16. Atom NZ from LYS33 in group 7 for the CDK2 dataset.

(aromatic/hydrophobic) from PHE80 are in group 1 but not in a pattern.

Out of the 27 atoms relevant for CDK2 interaction experimentally determined in Ref. 40, our strategy found 21, which represent about 78% of such atoms.

6.2.2. *Ricin*

In Ref. 17, authors co-crystalize ricin chain A with a transition state analogue inhibitor that mimics the sarcin-ricin recognition loop of the 28S rRNA. We consider the active site residues and atoms that the authors highlight as relevant in the interaction of ricin with subunit 28S rRNA as the experimentally determined patterns for the Ricin dataset. Next, we discuss some representative patterns.

GLY121:O (acceptor) is represented in the only GReMLIN pattern from group 4.

Atoms NH1 and NH2 (both donor) from ARG180 are depicted in the only pattern of group 7.

VAL81:O (acceptor) is present in the only pattern of group 4 and VAL81:N (donor) appears in the only pattern of group 9.

ASP96 OD1 (acceptor/negative) and OD2 (acceptor/negative) are depicted in group 10 only pattern.

ASP100:OD2 and ASP75:OD2 (both acceptor/negative) appear in the only pattern of group 21.

ASN78:ND2 (donor) is represented in the pattern of group 8.

Atoms CD1, CD2, CE1, CE2, CG and CZ from TYR80, all of them aromatic/hydrophobic, are depicted in the only pattern of group 2. Figure 17 shows these atoms from TYR80 in the graph that represents the interface between protein

Table 4. Binding site residues of CDK2 interacting with the 2 most potent sulfonamide analogue inhibitors.

Residue	Atom	GReMLIN
ASP145	CB	✓
	CG	✓
	OD1	✓
LYS33	CB	✓
	CD	✓
	CE	✓
	CG	✓
	NZ	✓
ASP86	N	✓
	CB	✓
	OD1	✓
	OD2	✓
LYS89	CB	✓
	CE	✓
	NZ	✓
GLN85	-	×
HIS84	O	✓
LEU83	N	✓
	O	✓
PHE82	CE2	•
	CZ	•
GLU81	O	✓
PHE80	CB	✓
	CG	•
	CD2	✓
	CE2	•
	CZ	•

Table 5. Active site residues of Ricin chain A interacting with a cyclic transition state analogue inhibitor

Residue	Atom	GReMLIN
GLY121	O	✓
ARG180	NH1	✓
	NH2	✓
VAL81	N	✓
	O	✓
ASP96	OD1	✓
	OD2	✓
ASP100	OD2	✓
ASP75	OD2	✓
ASN78	ND2	✓
TYR80	CD1	✓
	CD2	✓
	CE1	✓
	CE2	✓
	CG	✓
	CZ	✓
TYR123	N	✓
	CD2	✓
	CE2	✓
	CG	✓
GLU208*	-	×
GLU177	OE2	✓
ARG134*	-	×

✓ Residues/atoms found in patterns; • Found but not in patterns; × Not found.

1IFU:A and ligand FMC.

Atoms CD2, CE2 and CG from TYR123, all of them aromatic/hydrophobic, are present in the only pattern of group 1, while TYR123:N (donor) is represented in the pattern of group 16. Figure 18 shows one of the many occurrences of TYR123:N in group 16.

GLU177:OE2 (acceptor/negative) appears in the only pattern of group 19.

GLU208 and ARG134 do not directly interact with ligand according to Ref. 17.

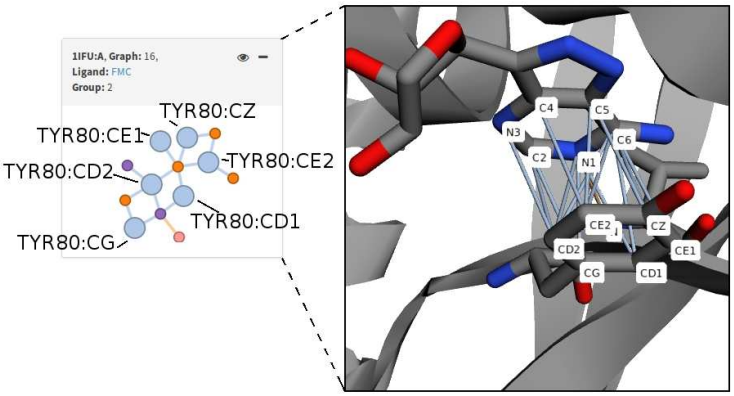


Fig. 17. Atoms CD1, CD2, CE1, CE2, CG and CZ from TYR80 in 2D graph and in a molecule viewer, graph16, in group 2 from the Ricin dataset.

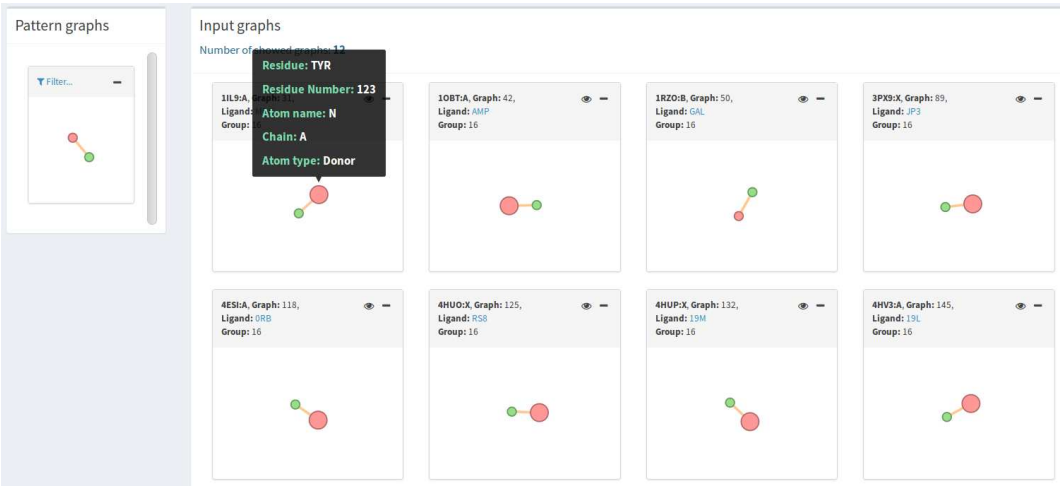


Fig. 18. Atom N from TYR123 in group 16 for the Ricin dataset.

Therefore it is expected that these residues do not appear in a pattern as GReMLIN does not consider water mediated interactions.

Out of the 23 atoms experimentally determined in Ref. 17 that are relevant for Ricin interaction with a transition state analogue inhibitor, our strategy found 21, which represent about 91%.

7. Conclusion

In this paper, we propose visGReMLIN, an interactive tool to visualize conserved interactions between protein and ligands for a set of related proteins. More specifically, we obtained a set of proteins from PDB, computed the interactions in the protein-ligand interface, and modeled such interactions as a bipartite graph at the atomic level. Each vertex represents an atom from protein or ligand and each edge depicts an interaction between a protein and a ligand atom. We labeled vertices and edges with physicochemical properties of atoms and interactions and used a strategy based on clustering analysis and frequent subgraph mining, which we proposed in Ref. 39, to compute conserved interactions in the protein-ligand interface. visGReMLIN delivers the input graphs and results of such strategy, allowing users to explore, filter, and understand conserved interaction patterns that are relevant for a variety of biological processes.

Our strategy is able to find 78% of the experimentally determined patterns for the CDK2 dataset and 91% of such patterns for the Ricin dataset in a totally automatic manner, using data available in PDB, without any manual support from domain specialists.

As future work, we intend to implement a more general version of our tool to permit users to choose their own dataset of interest to perform analysis of conserved patterns in the protein-ligand interface as well as to visualize and explore such patterns. Additionally, we want to perform a superposition of graphs taking as reference the conserved patterns, which will allow users to see at a glance which are the protein and ligand atoms involved in patterns. Finally, we plan to systematically collect user insights about the visGReMLIN to improve our visualization strategy considering the needs of domain specialists.

Funding

This work has been supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (grant number 477587/2013-5) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

References

1. Anand P, Nagarajan D, Mukherjee S, Chandra N, PLIC: protein-ligand interaction clusters, *Database (Oxford)* **2014**:bau029, 2014.
2. Balakin KV, Tkachenko SE, Lang SA, Okun I, Ivashchenko AA, Savchuk NP, Property-based design of gpcr-targeted library, *Journal of chemical information and computer sciences* **42**(6):1332–1342, 2002.
3. Berman HM, *et al.*, The Protein Data Bank, *Nucleic Acids Research* **28**(1):235–242, 2000. doi:10.1093/nar/28.1.235, URL `http://nar.oxfordjournals.org/content/28/1/235.abstract`.
4. Clark AM, Labute P, 2D depiction of protein-ligand complexes, *J Chem Inf Model* **47**(5):1933–1944, 2007.
5. Cordella LP, *et al.*, A (sub) graph isomorphism algorithm for matching large graphs,

IEEE transactions on pattern analysis and machine intelligence **26**(10):1367–1372, 2004.

6. Csardi G, Nepusz T, The igraph software package for complex network research, *InterJournal* **Complex Systems**:1695, 2006, URL `http://igraph.org`.
7. da Silveira CH, Pires DE, Minardi RC, Ribeiro C, Veloso CJ, Lopes JC, Meira W, Neshich G, Ramos CH, 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.
8. Danishuddin M, Khan AU, Structure based virtual screening to discover putative drug candidates: Necessary considerations and successful case studies, *Methods* **71**:135–145, 2015.
9. de Beer TA, Berka K, Thornton JM, Laskowski RA, PDBsum additions, *Nucleic Acids Res* **42**(Database issue):D292–296, 2014.
10. Desaphy J, Bret G, Rognan D, Kellenberger E, sc-PDB: a 3D-database of ligandable binding sites–10 years on, *Nucleic Acids Res* **43**(Database issue):399–404, 2015.
11. Durrant JD, McCammon JA, BINANA: a novel algorithm for ligand-binding characterization, *J Mol Graph Model* **29**(6):888–893, 2011.
12. Fuller JC, *et al.*, Ligdig: a web server for querying ligand–protein interactions, *Bioinformatics* **31**(7):1147–1149, 2015.
13. Gallina AM, Bisignano P, Bergamino M, Bordo D, PLI: a web-based tool for the comparison of protein-ligand interactions observed on PDB structures, *Bioinformatics* **29**(3):395–397, 2013.
14. Gonçalves-Almeida V, Pires DE, de Melo-Minardi RC, da Silveira CH, Meira W, Santoro MM, Hydropace: understanding and predicting cross-inhibition in serine proteases through hydrophobic patch centroids, *Bioinformatics* **28**(3):342–349, 2012.
15. Hanson RM, Prilusky J, Renjian Z, Nakane T, Sussman JL, Jsmol and the next-generation web-based representation of 3d molecular structure as applied to proteopedia, *Israel Journal of Chemistry* **53**(3-4):207–216, 2013.
16. Hendlich M, Bergner A, Gunther J, Klebe G, Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions, *J Mol Biol* **326**(2):607–620, 2003.
17. Ho MC, *et al.*, Transition state analogues in structures of ricin and saporin ribosome-inactivating proteins, *Proceedings of the National Academy of Sciences* **106**(48):20276–20281, 2009.
18. Huan J, *et al.*, Spin: mining maximal frequent subgraphs from graph databases, *Proceedings of the tenth ACM SIGKDD*, ACM, pp. 581–586, 2004.
19. Humphrey W, Dalke A, Schulten K, VMD: visual molecular dynamics, *J Mol Graph* **14**(1):33–38, 1996.
20. Jacob L, Vert JP, Protein-ligand interaction prediction: an improved chemogenomics approach, *Bioinformatics* **24**(19):2149–2156, 2008.
21. Kahraman A, *et al.*, Shape variation in protein binding pockets and their ligands, *J Mol Biol* **368**(1):283–301, 2007. doi:10.1016/j.jmb.2007.01.086, URL `http://dx.doi.org/10.1016/j.jmb.2007.01.086`.
22. Kasahara K, Kinoshita K, GIANT: pattern analysis of molecular interactions in 3D structures of protein-small ligand complexes, *BMC Bioinformatics* **15**:12, 2014.
23. Kaufman L, Rousseeuw P, *Clustering by means of medoids*, North-Holland, 1987.
24. Klabunde T, Chemogenomics approaches to ligand design, *Ligand Design for G Protein-coupled Receptors* pp. 115–135.
25. Koyutürk M, *et al.*, An efficient algorithm for detecting frequent subgraphs in biological networks, *Bioinformatics* **20**(suppl 1):i200–i207, 2004.

26. Laskowski RA, *et al.*, Ligplot+: multiple ligand–protein interaction diagrams for drug discovery, *Journal of chemical information and modeling* **51**(10):2778–2786, 2011.
27. Liu H, Sun J, Guan J, Zheng J, Zhou S, Improving compound–protein interaction prediction by building up highly credible negative samples, *Bioinformatics* **31**(12):i221–i229, 2015.
28. Mancini AL, , *et al.*, 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.
29. Mancini AL, Higa RH, Oliveira A, Dominiquini F, Kuser PR, Yamagishi ME, Togawa RC, 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.
30. Medina-Franco JL, Méndez-Lucio O, Martinez-Mayorga K, Chapter one-the interplay between molecular modeling and chemoinformatics to characterize protein–ligand and protein–protein interactions landscapes for drug discovery, *Advances in protein chemistry and structural biology* **96**:1–37, 2014.
31. Nakane T, *GLmol-Molecular Viewer on WebGL/Javascript, Version 0.47*, 2014.
32. Okabe A, *et al.*, *Spatial tessellations: concepts and applications of Voronoi diagrams*, Wiley, 2009.
33. Peason K, On lines and planes of closest fit to systems of point in space, *Philosophical Magazine* **2**:559–572, 1901.
34. Pires DE, *et al.*, acsm: noise-free graph-based signatures to large-scale receptor-based ligand prediction, *Bioinformatics* **29**(7):855–861, 2013.
35. Poupon A, Voronoi and voronoi-related tessellations in studies of protein structure and interaction, *Current opinion in structural biology* **14**(2):233–241, 2004.
36. Rognan D, Chemogenomic approaches to rational drug design, *British journal of pharmacology* **152**(1):38–52, 2007.
37. Rousseeuw PJ, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of computational and applied mathematics* **20**:53–65, 1987.
38. Salentin S, Schreiber S, Haupt VJ, Adasme MF, Schroeder M, PLIP: fully automated protein-ligand interaction profiler, *Nucleic Acids Res* **43**(W1):W443–447, 2015.
39. Santana CA, Cerqueira FR, da Silveira C H, Fassio AV, Melo-Minardi RC, Silveira SA, GReMLIN: a graph mining strategy to infer protein-ligand interaction patterns, *Bioinformatics and Bioengineering (BIBE), 2016 IEEE 16th International Conference on*, IEEE, 2016, *paper accepted*.
40. Schonbrunn E, *et al.*, Development of Highly Potent and Selective Diaminothiazole Inhibitors of Cyclin-Dependent Kinases, *Journal of Medicinal Chemistry* **56**(10):3768–3782, 2013.
41. Schreyer AM, Blundell TL, CREDO: a structural interactomics database for drug discovery, *Database (Oxford)* **2013**:bat049, 2013.
42. Schrödinger, *Schrödinger Release 2016-3: Maestro, version 10.7*, 2016, https://www.schrodinger.com/.
43. Silveira SA, *et al.*, Revealing protein-ligand interaction patterns through frequent subgraph mining, *Proceedings of BIOCOMP*, The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), p. 50, 2015.
44. Sobolev V, *et al.*, Automated analysis of interatomic contacts in proteins, *Bioinformatics* **15**(4):327–332, 1999.
45. Stierand K, Rarey M, Drawing the PDB: Protein-Ligand Complexes in Two Dimensions, *ACS Med Chem Lett* **1**(9):540–545, 2010.

46. Velankar S, van Ginkel G, Alhroub Y, Battle GM, Berrisford JM, Conroy MJ, Dana JM, Gore SP, Gutmanas A, Haslam P, Hendrickx PM, Lagerstedt I, Mir S, Fernandez Montecelo MA, Mukhopadhyay A, Oldfield TJ, Patwardhan A, Sanz-Garcia E, Sen S, Slowley RA, Wainwright ME, Deshpande MS, Iudin A, Sahni G, Salavert Torres J, Hirshberg M, Mak L, Nadzirin N, Armstrong DR, Clark AR, Smart OS, Korir PK, Kleywegt GJ, PDBe: improved accessibility of macromolecular structure data from PDB and EMDB, *Nucleic Acids Res* **44**(D1):D385–395, 2016.
47. Wallace AC, Laskowski RA, Thornton JM, LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions, *Protein Eng* **8**(2):127–134, 1995.
48. Yamaguchi A, Iida K, Matsui N, Tomoda S, Yura K, Go M, Het-PDB Navi.: a database for protein-small molecule interactions, *J Biochem* **135**(1):79–84, 2004.
49. Yan X, Han J, gspan: Graph-based substructure pattern mining, *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, IEEE, pp. 721–724, 2002.

Alexandre V. Fassio received his Bachelor of Computer Science from Universidade Federal de Lavras, Brazil, in 2013. In 2015, he received his M.Sc. degree in Bioinformatics from Universidade Federal de Minas Gerais, Brazil, and, since then, he is pursuing his PhD in Bioinformatics at the same university, working on protein-ligand interaction.

Charles A. Santana received his Bachelor of Computer Science from Universidade Estadual do Sudoeste da Bahia, in 2014, and he is pursuing his M.Sc. from Universidade Federal de Viçosa, both in Brazil. He is interested in machine learning, optimization and graph pattern mining applied to Bioinformatics.

Fabio R. Cerqueira graduated in Computer Science at Universidade Federal de Vicosa, in 1996, and received his M.Sc. degree also in Computer Science from Unicamp, in 1999, both in Brazil. He received his PhD in Biomedical Informatics from University for Health Sciences, Medical Informatics and Technology (UMIT), Austria, in 2010. He is professor at Universidade Federal de Viosa and his recent projects are focused on short RNAs and short ORFs.

João P. R. Romanelli received his bachelor in Mathematics in 2004 and his M.Sc. degree also in Mathematics in 2006, both from Universidade Federal de Minas Gerais. He received his D.Sc. of degre in Mathematics in 2011 from Pontifícia Universidade Católica do Rio de Janeiro and he is currently a professor at Universidade Federal de Itajubá. His recent research interests are graph theory, matrix decomposition and functional density theory.

Carlos H. da Silveira graduated in Computer Science at Universidade Federal de Minas Gerais, in 1996, and received his D.Sc. degree in Bioinformatics at the same university. He is currently a professor at Universidade Federal de Itajubá. His research interests are structural bioinformatics, computational modeling of

biomolecules, database integration and sequence-structure relationship.

Raquel C. de Melo-Minardi received her bachelor of Computer Science, in 2004, and D.Sc. degree in Bioinformatics, in 2008, both from Universidade Federal de Minas Gerais. In 2008, she joined Comissariat l'Energie Atomique et aux nergies Alternatives / Genoscope, France, for a 1-year postdoc. She is currently a professor at Universidade Federal de Minas Gerais. Her main research interests are bioinformatics and data visualization.

Sabrina de A. Silveira received her bachelor of Computer Science, in 2008, from Universidade Federal de Minas Gerais and her D.Sc. degree in Bioinformatics, in 2013, from the same university. She is currently professor at Universidade Federal de Viçosa. Her research interests include structural bioinformatics, data mining and data visualization.