UNIVERSIDADE FEDERAL DE VIÇOSA

ppiGremlin - Uma estratégia para prospecção de padrões em interfaces proteína-proteína baseada em mineração de grafos

Felippe Cathoud de Queiroz
Magister Scientiae

VIÇOSA - MINAS GERAIS
2019

FELIPPE CATHOUD DE QUEIROZ

ppiGremlin - Uma estratégia para prospecção de padrões em interfaces proteína-proteína baseada em mineração de grafos

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*.

Orientadora: Sabrina de A. Silveira

Coorientadores: Maria G. de A. Oliveira
Adriana M. P. Vargas
Giovanni V. Comarela

VIÇOSA - MINAS GERAIS
2019

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
Q3p
2019

Queiroz, Felippe Cathoud de, 1990-
ppiGremlin - uma estratégia para prospecção de padrões em interfaces proteína -proteína baseada em mineração de grafos / Felippe Cathoud de Queiroz. – Viçosa, MG, 2019.
1 dissertação eletrônica (61 f.): il. (algumas color.).

Orientador: Sabrina de Azevedo Silveira.
Dissertação (mestrado) - Universidade Federal de Viçosa, Departamento de Informática, 2019.
Referências bibliográficas: f. 56-61.
DOI: https://doi.org/10.47328/ufvbbt.2024.673
Modo de acesso: World Wide Web.

1. Interação proteína-proteína - Processamento de dados. 2. Mineração de dados (Comutação). I. Silveira, Sabrina de Azevedo, 1983-. II. Universidade Federal de Viçosa. Departamento de Informática. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDD 22. ed. 572.640285

Bibliotecário(a) responsável: Bruna Silva CRB-6/2552

FELIPPE CATHOUD DE QUEIROZ

ppiGremlin - Uma estratégia para prospecção de padrões em interfaces proteína-proteína baseada em mineração de grafos

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: 22 de novembro de 2019.

Assentimento:

_____

Felippe Cathoud de Queiroz
Autor

_____

Sabrina de Azevedo Silveira
Orientadora

À minha família e aos meus amigos, que me ofereceram suporte necessário para concluir este trabalho.

AGRADECIMENTOS

Agradeço primeiramente a Deus por me permitir esta jornada desafiadora e por ter me sustentado em todos os momentos.

Agradeço à minha família pelo apoio incondicional que me proporcionou a tranquilidade e o conforto necessários para vencer esta etapa, e pelo incentivo em sempre prosseguir.

Aos meus amigos Camila e Nildo, pelos momentos que compartilhamos em Viçosa, pela paciência e pelo suporte nos momentos difíceis.

Aos professores que contribuíram para a conclusão dessa trajetória, especialmente à professora Sabrina de Azevedo Silveira. Agradeço à sua imensa dedicação e paciência, e por me auxiliar de maneira tão inteligente na solução dos problemas.

À Universidade Federal de Viçosa e ao Departamento de Informática, pela sólida formação acadêmica que me foi oferecida ao longo dos anos em que estive em Viçosa.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

E finalmente, a todos aqueles que, direta ou indiretamente, contribuíram de alguma forma para a execução desse trabalho, os meus sinceros agradecimentos.

"... e sem Ele nada do que foi feito se fez"
João 1:3

RESUMO

QUEIROZ, Felippe Cathoud de, M.Sc., Universidade Federal de Viçosa, novembro de 2019. **ppiGremlin - Uma estratégia para prospecção de padrões em interfaces proteína-proteína baseada em mineração de grafos** . Orientadora: Sabrina de Azevedo Silveira. Coorientadores: Maria Goreti de Almeida Oliveira, Adriana Maria Patarroyo Vargas e Giovanni Ventorim Comarela.

Interações proteína-proteína (PPI's) estão presentes em praticamente todos os processos biológicos nos organismos vivos, desde a transcrição gênica até os mais altos níveis da sua organização molecular e estrutural. Compreender os mecanismos moleculares envolvidos nessas interações apresenta um contribuição significativa para o desenvolvimento de fármacos, projeto de peptídeos e identificação de alvos moleculares. Neste sentido, este trabalho propõe ppiGReMLIN, uma estratégia baseada em mineração de grafos, para inferir padrões de interação na interface de ligação entre proteínas. Buscamos, através de uma representação simplificada das interfaces, encontrar padrões que reflitam estruturas conservadas em PPI's reais. Para isso, as interfaces de interação são modeladas como grafos bipartidos, cujos vértices e arestas representam, respectivamente, átomos e interações não-covalentes entre eles. Interações são computadas através de análise de contatos, segundo um critério de distância, de acordo com as propriedades físico-químicas dos átomos. Em sequência, o conjunto de grafos produzidos é fragmentado em grupos de acordo com as suas similaridades, utilizando um algoritmo de agrupamento, sobre os quais, em seguida é aplicada mineração de subgrafos frequentes, visando extrair padrões relevantes em cada grupo. As estratégia foi verificada mediante dados estruturais de complexos proteína-proteína extraidos do *Protein Data Bank*. Duas bases de dados com relevância biológica foram utilizadas na avaliação da estratégia: BCL-2/BH3 e Serino-Proteases. ppiGReMLIN foi capaz de identificar padrões de interação frequentes em ambas as bases de dados, respaldados por estudos experimentais de acordo com a literatura.

Palavras-chave: interações proteína-proteína, mineração de dados, padrões de iteração

ABSTRACT

QUEIROZ, Felippe Cathoud de, M.Sc., Universidade Federal de Viçosa, November, 2019. **ppiGremlin - A strategy for pattern prospecting at protein-protein interfaces based on graph mining**. Adviser: Sabrina de Azevedo Silveira. Co-advisers: Maria Goreti de Almeida Oliveira, Adriana Maria Patarroyo Vargas and Giovanni Ventorim Comarela.

Protein-protein interactions are prevalent in practically all biological processes in living organisms, from gene transcription to the highest levels of molecular and structural organization. A better understanding of the molecular mechanisms involved in such interactions may offer significant improvement in drug development, peptide design, identification of molecular targets and many others areas. For this purpose, this work proposes ppiGReMLIN, a subgraph mining based strategy to infer interaction patterns in protein-protein interfaces. By means of a simplified representation of interfaces, we aim to uncover relevant patterns that illustrate conserved structural arrangements in real PPI's. Our strategy models protein-protein interfaces as bipartite graphs, whose vertices represent protein atoms, and edges represent interactions among them. Nodes and edges are labeled according to its physicochemical properties and a distance criteria. Then, a clustering analysis is performed on the set of graphs to characterize them according to their similarities, and a subgraph mining task is ensued in order to extract relevant patterns over each group. The strategy was verified against structural data of protein-protein complexes harnessed from the *Protein Data Bank*. Two relevant datasets were used in the evaluation: BCL-2/BH3 e Serino Protease. ppiGReMLIN was able to disclose relevant interaction patterns in both datasets, confirmed by experimental studies according to literature.

Keywords: protein-protein interactions, data mining, interaction patterns

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

Capítulo 1

Referencial Teórico

1.1 Proteínas

Proteínas são as macromoléculas mais abundantes em organismos vivos, ocorrendo em todas as células, assim como em cada um dos componentes celulares. Elas são encontradas na composição de enzimas, anticorpos, hormônios, antibióticos, toxinas, fibras musculares e etc. Proteínas desempenham funções cruciais em diversos processos biológicos como catálise de reações, transporte e armazenamento de substâncias, transmissão de impulsos nervosos, controle de crescimento, diferenciação celular e muitos outros (Nelson et al., 2008).

Os aminoácidos são as unidades básicas na formação das proteínas. Eles possuem uma estrutura genérica (figura 1.1) composta por um grupo carboxila, um grupo amino e uma cadeia lateral (R) ligados a um mesmo átomo de carbono (carbono α). As cadeias laterais diferem entre si na sua composição, tamanho e estrutura, possuindo propriedades físico-químicas diferentes, como carga elétrica, polaridade e aromaticidade. Dentre os diversos tipos de aminoácidos existentes na natureza, apenas 20 são comumente utilizados na construção das proteínas (figura 1.2).

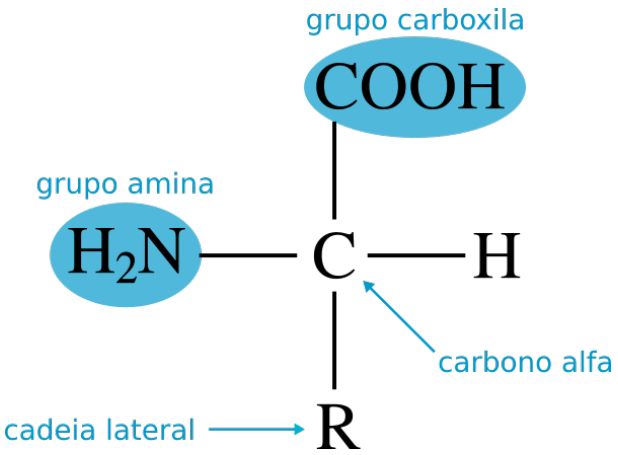


Figura 1.1: Estrutura de um aminoácido.

Os aminoácidos são conectados sequencialmente por meio de ligações petídicas para formar cadeias polipeptídicas. Estas são ligações covalentes entre um átomo de

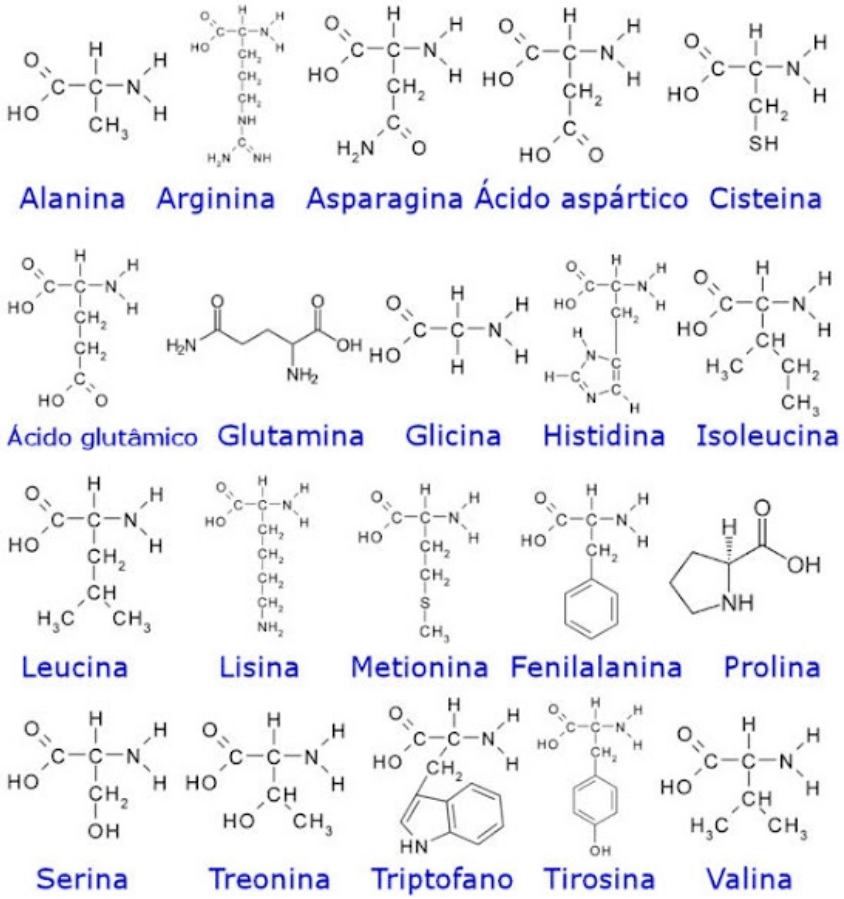


Figura 1.2: Tipos de aminoácidos.

carbono e nitrogênio resultantes da reação entre os grupos carboxila e amino respectivamente de dois aminoácidos diferentes com consequente perda de uma molécula de água. Nesse processo, à parte restante de cada aminoácido que compõe a cadeia da-se o nome de resíduo de aminoácido ou simplesmente resíduo. Cada proteína possui uma sequência única, que se caracteriza pelo número de aminoácidos que a compõem, bem como, os tipos de cada aminoácido da sequência.

A estrutura das proteínas pode ser descrita em quatro níveis de organização segundo a sua complexidade. O nível mais básico compreende a sequência dos aminoácidos conectados por ligações covalentes e é chamado de estrutura primária. A estrutura secundária é composta principalmente pelo arranjo de trechos da sequencia de aminoácidos em padrões estruturais mais estáveis, tais como α-hélices e folhas β. O enovelamento completo da cadeia principal formam estruturas tridimensionais ainda mais organizadas que constituem a estrutura terciária das proteínas. A partir daí, quando uma proteína é composta de múltiplas cadeias polipeptídicas, à sua estrutura designa-se o nome de estrutura quaternária (Alberts, 2017).

Sobre o mecanismo de enovelamento, sabe-se que as estruturas resultantes desse processo são predominantemente produto de um conjunto interações fracas que agem

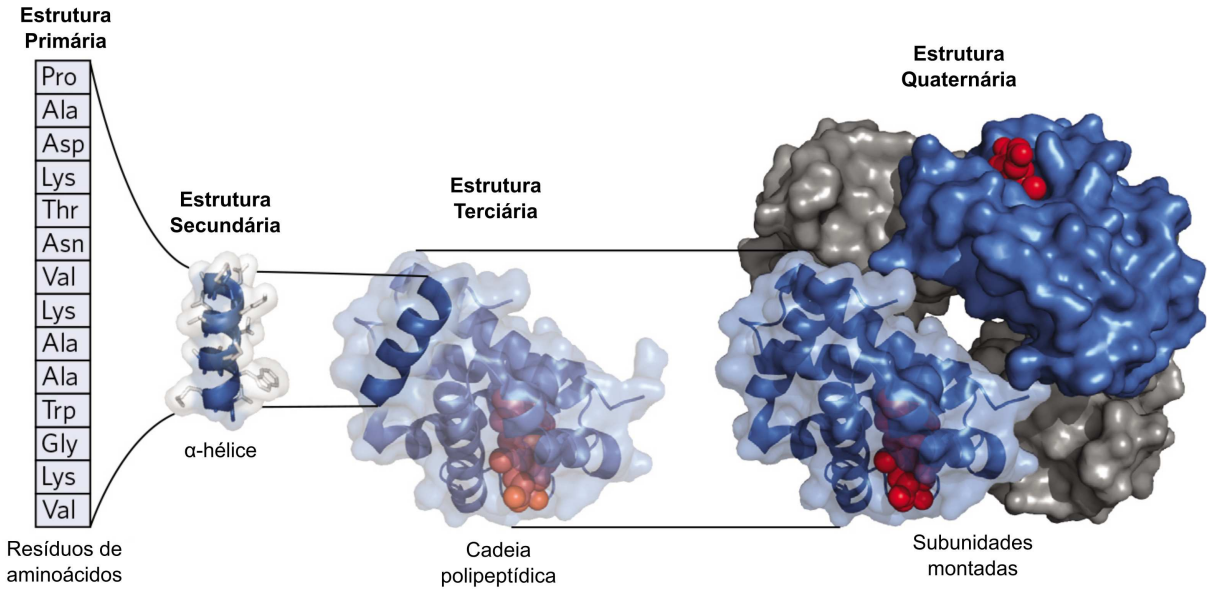


Figura 1.3: Estrutura de uma proteína.(Fonte: Adaptado de (Nelson and Cox, 2018))

internamente entre os aminoácidos das cadeias polipeptídicas e entre estes aminoácidos e o ambiente externo ou solvente. Estas mesmas interações influenciam de modo significativo o processo de construção de complexos proteicos e as interações entre proteínas (Ofran and Rost, 2003). Nesse âmbito, as cadeias se dobram assumindo conformações energeticamente mais favoráveis, o que se traduz em estruturas mais estáveis. Sabe-se também que a sequência dos aminoácidos é determinante na estrutura da proteína e que proteínas diferentes possuem sequências diferentes (Fromm and Hargrove, 2011).

A função de uma proteína está diretamente relacionada à sua estrutura, determinando como ela interage com outras moléculas. Em geral, essa estrutura não é estática e sofre mudanças em sua conformação mediante a interação, que pode ou não ser reversível. No caso das interações reversíveis, as moléculas com as quais as proteínas interagem recebem o nome de ligantes, que podem ser de diversos tipos, mesmo outras proteínas (Nelson et al., 2008). No escopo deste trabalho consideramos apenas proteínas ou pequenas cadeias polipeptídicas como ligantes.

1.2 Interações Proteína-Proteína

Interações entre proteínas ou interações proteína-proteína (PPI) estão presentes na maioria dos processos biológicos nos organismos. Praticamente toda função biológica envolve comunicação entre proteínas, e num sentido mais amplo, elas constituem redes de interações extremamente complexas que determinam o comportamento dos sistemas biológicos (Nevola and Giralt, 2015).

PPI's podem ser classificadas de diversas maneiras segundo a sua composição,

estrutura e afinidade (Scott et al., 2016). Com relação à sua composição, quando são formadas por subunidades iguais, elas são chamadas homéricas, e heteroméricas quando suas subunidades são diferentes . Pelo critério de afinidade, PPI's obrigatórias são interações mais fortes e de maior duração quando comparadas às não-obrigatórias ou transientes. De acordo com estrutura, considera-se a estrutura individual de cada uma das proteínas componentes (se é um proteína globular ou um cadeia peptídica) e as suas possíveis mudanças de conformação:

- Pares de proteínas globulares que interagem com epítopo não continuo, sem mudança de conformação (figura 1.4a);
- Pares de proteínas globulares onde uma ou ambas sofrem alteração em sua conformação (figura 1.4b);
- Interação entre uma proteína globular e uma cadeia peptídica isolada (figuras 1.4c e 1.4d);
- Interação entre duas cadeias peptídicas isoladas (figura 1.4e).

Os dois últimos podem ainda ser subdivididos considerando as mudanças de conformação das cadeias peptídicas. Alguns peptídeos ou trechos de cadeia na superfície de algumas proteínas apresentam estrutura nativa desordenada, e alteram sua conformação apenas mediante interação.

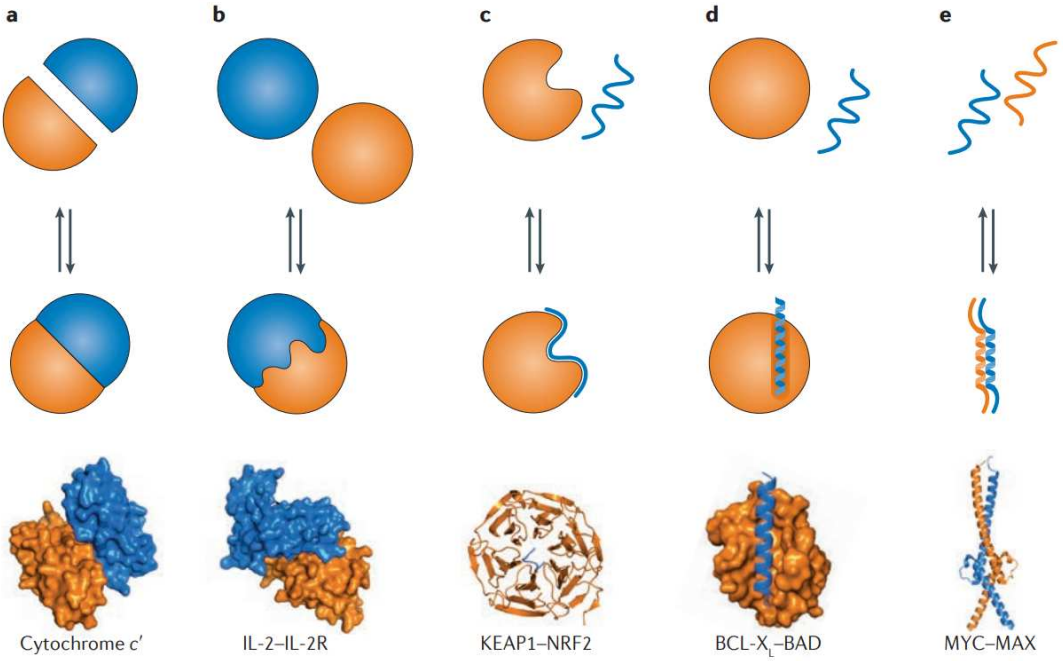


Figura 1.4: Estruturas de PPI's. (Fonte: (Scott et al., 2016))

Uma avanço importante no entendimento das PPI's, foram os *hot spots* (Moreira et al., 2007). Estes são resíduos ou regiões na interface de interação de complexos

proteína-proteína que contribuem de forma mais acentuada para sua energia de ligação. Experimentalmente, os *hot spots* podem ser caracterizados através de um processo conhecido como *Alanine Scanning*, que consiste na análise da energia de interação mediante a substituição de resíduos alvo na sequência da proteína por resíduos de glicina ou alanina.

Existe grande interesse no estudo de PPI's e na sua modulação, no sentido de compreender melhor os mecanismos biológicos dentro dos organismos, afim de elucidar os mecanismos responsáveis por doenças, e em última instância, desenvolver novos terapêuticos ou estratégias de tratamento das mesmas (Nevola and Giralt, 2015).

No entanto, o desenvolvimento de moduladores moleculares de PPI's é desafiador quando comparado aos alvos convencionais, representados por interações entre proteínas e pequenos ligantes. As proteínas que compõem essas interfaces de interação em geral possuem área superficial relativamente pequena ($\sim 300 - 1000$ Å^2) com a presença de sulcos ou bolsões de pequeno volume. Em contrapartida, PPI's apresentam-se em classes estruturais variadas, como mostrado anteriormente, algumas com características bem distintas. As interfaces de contato em algumas proteínas globulares, por exemplo, são mais planas, e apresentam área superficial maior ($\sim 1500 - 3000$ Å^2) (Nevola and Giralt, 2015).

1.3 Grafos

Grafos são modelos de representação de dados com aplicações em diversas áreas, como modelagem de redes de transporte e de comunicação em pesquisa operacional, modelagem de problemas de engenharia e ciência econômica em teoria dos jogos, estudo de moléculas em biologia e química, e muitos outros (Shirinivas et al., 2010).

Um grafo G é um par ordenado $(V(G), E(G))$, formado por um conjunto $V(G)$ de vértices e um conjunto $E(G)$, disjunto em relação a $V(G)$, de arestas, junto com uma função de incidência ψ_G, que associa a cada aresta de G um par de vértices de G (figura 1.5). Considerando que a é uma aresta, e u e v são vértices, tais que $\psi_G(a) = \{u, v\}$, então diz-se que a conecta u e v, além disso os vértices u e v são *extremidades* de a, e u e v são *adjacentes* (Bondy and Murty, 2011).

Um grafo pode conter laços (arestas com ambas as extremidades em um mesmo vértice) ou arestas paralelas. Na figura 1.5, a aresta b representa um laço e as arestas d e f são arestas paralelas, conectando o mesmo par de vértices (x, w). A este tipo de estrutura dá-se o nome de multigrafo ou pseudógrafo, embora exista divergência entre autores, e alguns considerem como multigrafos apenas aqueles com arestas paralelas (Harary, 1969). Neste texto, o termo multigrafo aplica-se somente a este último tipo. Um grafo que não apresenta laços ou arestas paralelas é chamado de grafo simples.

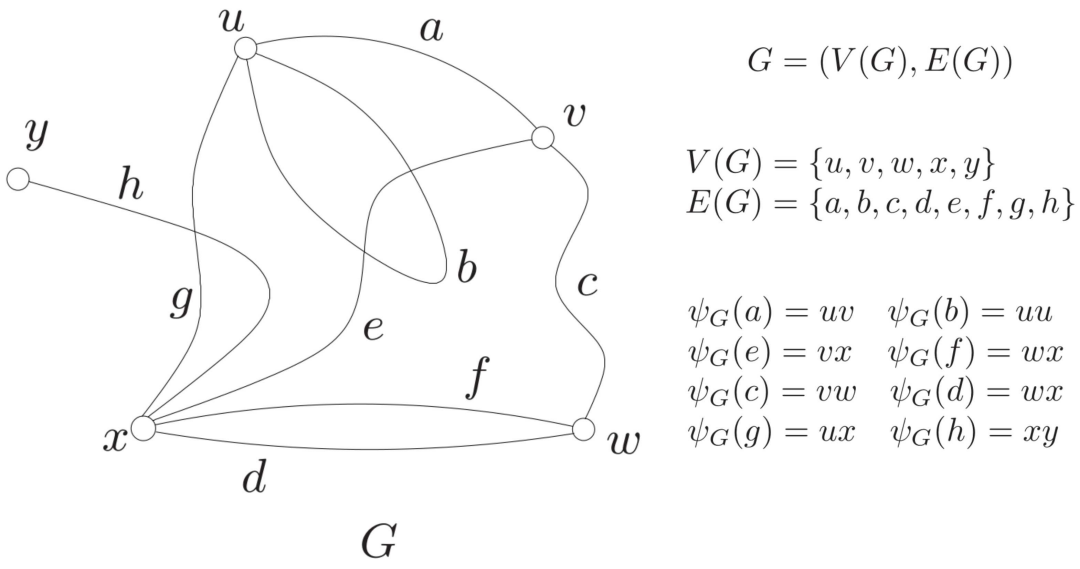


Figura 1.5: Exemplo de estrutura de um grafo. (Fonte: (Bondy and Murty, 2011))

Um grafo F é subgrafo de um grafo G, se $V(F) \subseteq V(G)$, $E(F) \subseteq E(G)$, e ψ_F é a restrição de ψ_G sobre $E(F)$. Nesse caso denota-se $G \supseteq F$ ou $F \subseteq G$, e diz-se, respectivamente, que G contém F ou F esta contido em G.

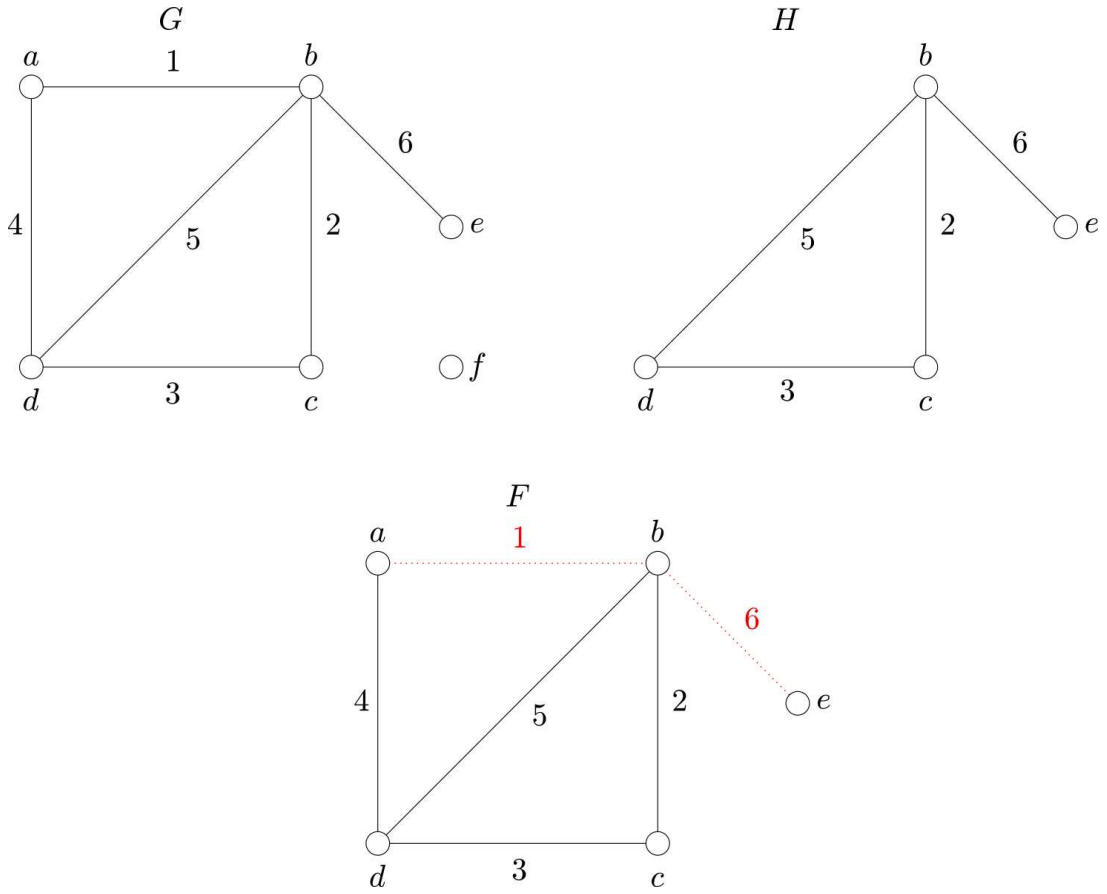


Figura 1.6: Exemplo de grafos e subgrafos induzidos

Seja Y um subconjunto dos vértices de um grafo G, isto é, $Y \subseteq V(G)$. Então $H = G[Y]$ é o grafo que tem Y como seu conjunto de vértices, e cujo conjunto de arestas é composto pelas arestas de G que possuem ambas extremidades em Y (H na Figura 1.6). Analogamente, se S é um subconjunto das arestas de um grafo G, então $F = G[S]$ é o grafo que tem S como seu conjunto de arestas, e cujo conjunto de vértices é composto pelos vértices nas extremidades das arestas de S (F na figura 1.6). H e F são denominados subgrafos de G induzidos por Y e S respectivamente (Bondy and Murty, 2011).

Um grafo G é dito conexo ou conectado se para toda partição do seu conjunto de vértices em dois conjuntos X e Y não vazios, existe uma aresta com uma extremidade em X e outra em Y; se isso não acontece, G é dito desconexo ou desconectado. Além disso, cada subgrafo maximal conectado de um grafo G é chamado de componente conectado de G ou simplesmente componente de G (Figura 1.6) (Harary, 1969).

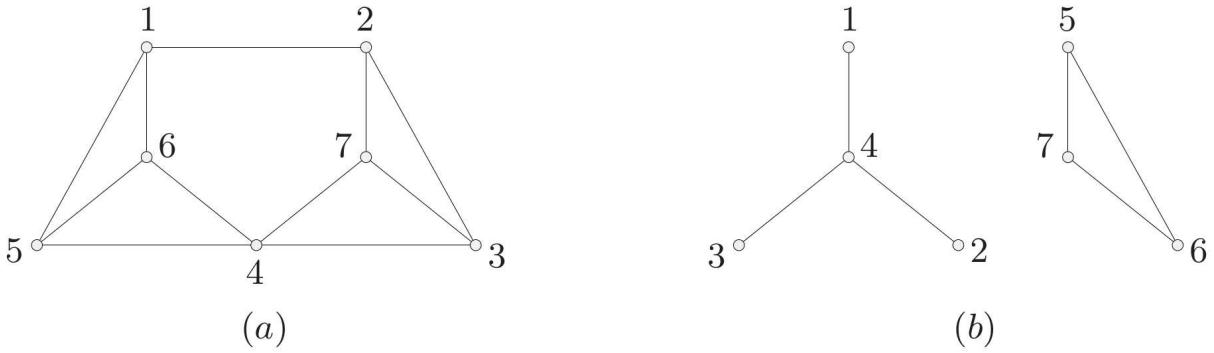


Figura 1.7: Exemplos de grafos conexos (a) e desconexos(b)

Um grafo linha $L(G)$ de um grafo G é o grafo cujo conjunto de vértices é formado pelas arestas de G, e onde tais vértices são conectados apenas se as respectivas arestas em G possuem um vértice em comum. A figura 1.8 mostra um grafo G e o respectivo grafo linha $L(G)$ gerado (Harary, 1969).

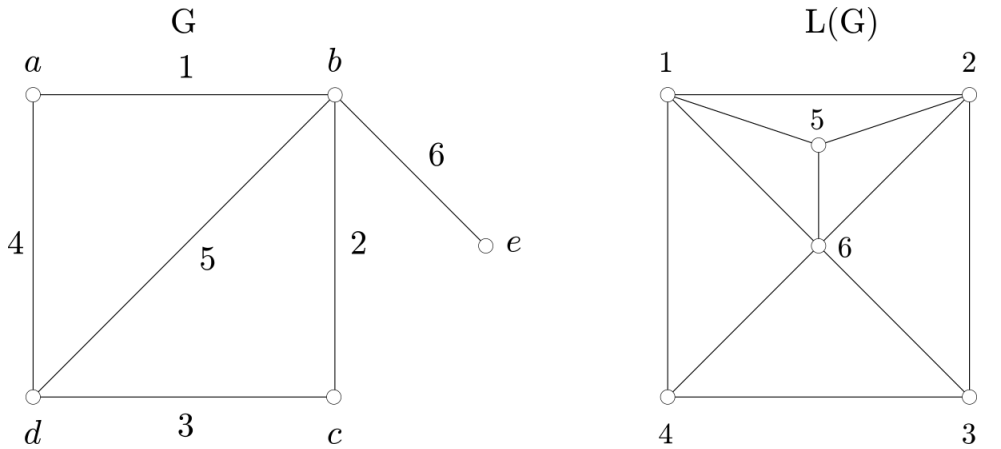
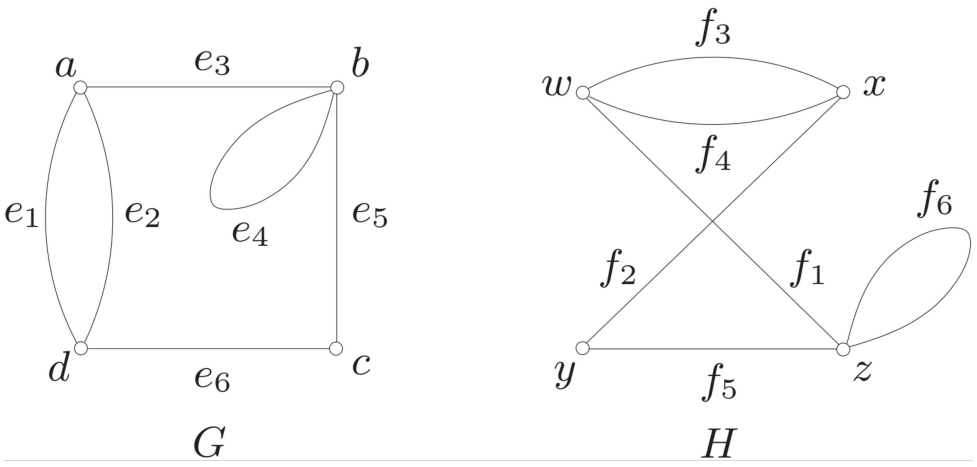


Figura 1.8: Exemplo de grafo linha *L(G)* gerado a a partir de um grafo *G*

Dois grafos G e H são ditos isomórficos ($G \cong H$) se existe bijeções $\theta : V(G) \to V(H)$ e $\phi : E(G) \to E(H)$, tal que $\psi_G(e) = uv$, se e somente se, $\psi_H(\phi(e)) = \theta(u)\theta(v)$ (Bondy and Murty, 2011). A figura 1.9 mostra um exemplo de grafos isomórficos e as suas respectivas bijeções. Para grafos rotulados, pode-se considerar ainda os rótulos dos vértices e arestas como parte do isomorfismo. Nesse caso, se l é a função que mapeia os rótulos para cada vértice e aresta do grafo, então adiciona-se as seguintes restrições à definição acima: $l_G(u) = l_H(\theta(u))$ e $l_G(uv) = l_H(\phi(uv))$ (Hsieh et al., 2006).



$$\theta := \begin{pmatrix} a & b & c & d \\ w & z & y & x \end{pmatrix} \qquad \phi := \begin{pmatrix} e_1 & e_2 & e_3 & e_4 & e_5 & e_6 \\ f_3 & f_4 & f_1 & f_6 & f_5 & f_2 \end{pmatrix}$$

Figura 1.9: Exemplo de grafos isomórficos. (Fonte: (Bondy and Murty, 2011))

Em biologia, grafos podem ser aplicados para a representação de estrutura de diversas moléculas assim como as suas interações. Saidi et al. (2009) mostram exemplos de modelagem de proteínas sob perspectivas diferentes: do nível de estrutura secundária, de resíduos e de átomos, onde a maior preocupação consiste em produzir um modelo coerente com a conformação estrutural real das estruturas.

1.4 Mineração de subgrafos

O principal objetivo da mineração de dados é extrair conhecimento estatisticamente significativo e útil a partir de um conjunto de dados (Han et al., 2011). No caso de dados representados na forma de grafos, a essa tarefa dá-se o nome de mineração de grafos.

Mineração de subgrafos frequentes (MSF), como parte essencial da mineração de grafos, tem como principal objetivo identificar subgrafos frequentes em um dado conjunto, cujo número de ocorrências no conjunto satisfaça um valor mínimo especificado. As técnicas existentes para lidar com esse problema podem ser classificadas

segundo o tipo do conjunto de dados como técnicas baseadas em grafo único e técnicas baseadas em transações. No primeiro caso, a base de dados é composta de um único grafo, geralmente grande, e as ocorrências de subgrafos são obtidas a partir de diferentes regiões desse único grafo. No outro caso, a base de dados é formada por um conjunto de grafos (transações), relativamente menores, e as ocorrências de subgrafos são obtidas nos diferentes grafos do conjunto. Ainda para os conjuntos de entrada baseados em transações, a contagem para a definição de suporte pode ser feita levando-se em conta as múltiplas ocorrências de um mesmo subgrafo em cada grafo do conjunto de entrada, ou simplesmente, considerando apenas o número de grafos de entrada onde há pelo menos uma ocorrência dos subgrafos (Jiang et al., 2013). Está última abordagem foi a abordagem utilizada no escopo deste trabalho, e é discutida abaixo.

Formalmente, considerando-se uma base de dados $\Omega = \{G_1, G_2, ..., G_T\}$, composta de um conjunto de transações, um grafo g, e um valor limite de suporte σ ($0 < \sigma < 1$), defini-se o conjunto de transações onde g é subgrafo como $\delta_\Omega(g) = \{G_i | g \subseteq G_i\}$, e o suporte de g como $sup_\Omega(g) = |\delta_\Omega(g)|/T$, onde $|\delta_\Omega(g)|$ é a cardinalidade de $\delta_\Omega(g)$, e T é o número de grafos (transações) em Ω. A figura 1.10 mostra um exemplo de um conjunto de grafos e alguns subgrafos frequentes encontrado com os respectivos suportes.

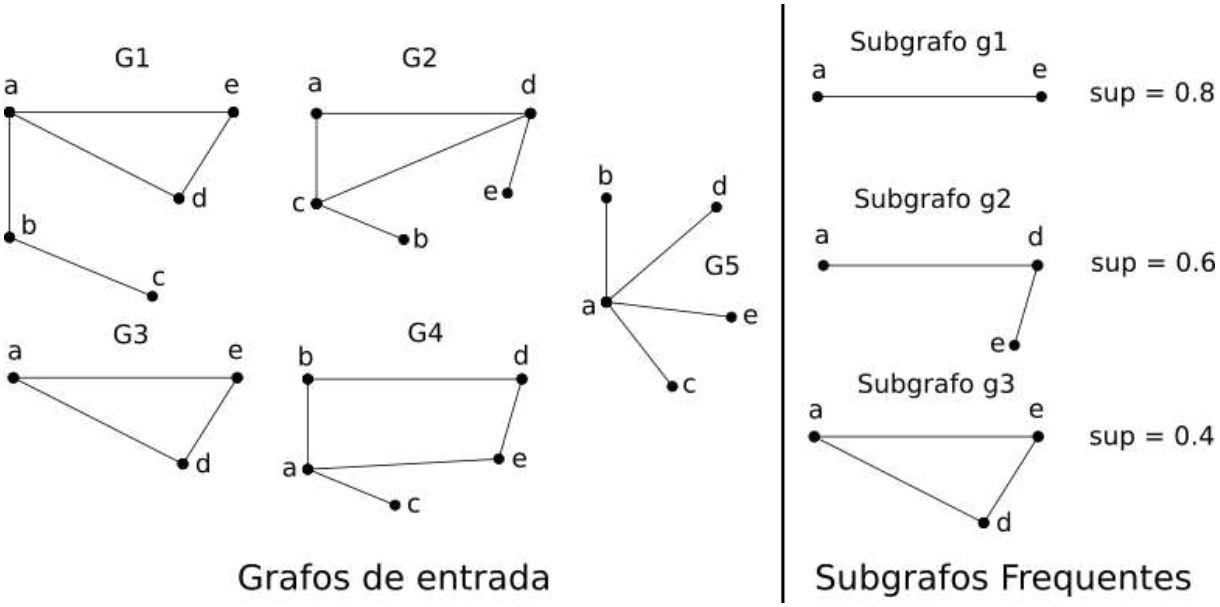


Figura 1.10: Subgrafos frequentes.

De maneira geral, a tarefa de mineração de subgrafos pode ser resumida em dois momentos — geração de candidatos e determinação da frequência de ocorrências, e é nestes momentos onde a maior parte dos esforços de pesquisa são direcionados, de modo a realiza-los de maneira eficiente. Para a geração de candidatos, busca-se evitar a geração de candidatos duplicados ou supérfluos. No caso da determinação

da frequência de ocorrências, a principal operação é a comparação de candidatos com subgrafos existentes nos grafos de entrada, o que é feito por meio de técnicas de isomorfismo.

Alguns exemplos de algoritmos de mineração de subgrafos frenquentes são: FSG (Kuramochi and Karypis, 2004), MoFa (Borgelt and Berthold, 2002) e gSpan (Yan and Han, 2002). Entre esses, o gSpan é um dos algoritmos mais conhecidos descritos na literatura ((Mrzic et al., 2018; Jiang et al., 2013)), e experimentalmente apresentou performance superior a uma ordem de magnitude em relação ao FSG (Yan and Han, 2002).

Um dos problemas da mineração de grafos é o número de padrões produzidos, que cresce exponencialmente em relação à base de dados de grafos minerados (Huan et al., 2004). Uma maneira de contornar esse problema, sem perda de informação, é através da utilização de subgrafos frequentes maximais. Os grafos frequentes maximais de um conjunto são aqueles que são frequentes no conjunto e que não são subgrafos de nenhum outro grafo frequente no conjunto. Segundo Koyutürk et al. (2004), em redes biológicas, os grafos maximais são mais interessantes para análise.

1.5 Agrupamento

Agrupamento ou *clustering* refere-se ao processo de examinar um conjunto de dados e dividi-lo em grupos, de modo que objetos em um mesmo grupo sejam similares entre si e objetos em grupos diferentes sejam distintos. Trata-se de uma modalidade exploratória de aṅalise de dados, expresso através um conjunto de técnicas e algoritmos de aprendizado não supervisionado. Nesse sentido, o principal objetivo é determinar um agrupamento inerente ao conjunto de dados considerando apenas os seus próprios atributos.

Agrupamento é um componente essencial em diversas aplicações dentro de análise de dados e aprendizado de máquina, como regressão, predição e mineração de dados (Han et al., 2006). Em áreas, como biologia, possui aplicação em análise de informações clínicas, filogenia, proteômica, genômica, análise de bases de dados biológicas, entre outras.

O métodos de agrupamento podem ser categorizados segundo a sua estratégia na abordagem dos dados. Zaki and Wagner Meira (2014) caracterizam os métodos em quatro categorias principais:

- **Métodos representativos**: dado o número de grupos k, estes métodos buscam dividir um conjunto de dados em k grupos onde cada grupo possui um elemento central que representa todo conjunto. A similaridade dos elementos é determinado em referencia a estes elementos representantes, geralmente segundo al-

guma distância entre os elementos. Exemplos: *K-means* , *K-medoids*, *Expectation Maximization* (Aggarwal, 2015);

- **Métodos hierárquicos**: o principal objetivo destes métodos é criar uma sequência de partições do conjunto de dados, que podem ser visualizados na forma de árvore ou dendograma de partições. As principais abordagens são agrupamento aglomerativo ou por divisão. No primeiro caso, inicia-se com cada elemento do conjunto de dados em grupos isolados, que são fusionados para formar grupos maiores de maneira progressiva, até que um número de grupos seja atingido, de acordo algum critério predefinido. No outro caso, a abordagem é inversa, e inicia-se com um único grupo contendo todos os elementos, que é particionado progressivamente.

- **Métodos baseados em densidade**: estes métodos utilizam a densidade local dos pontos como base para definição de grupos. São capazes de identificar regiões de alta densidade separadas por outras regiões com baixa densidade. Em relação as métodos representativos, eles são capazes de identificar grupos com tamanhos e formatos variados. No entanto, não são adequados para dados com grande variação de densidade. Exemplos: *DBSCAN (Ester et al., 1996)*, *DENCLUE (Hinneburg et al., 1998)*

- **Métodos baseados em grafos**: para este tipo de método, os elementos de dados são convertidos em vértices e a similaridade entre estes é representada pelos pesos das arestas que conectam os respectivos vértices. Em relação aos métodos anteriores, estes são capazes de identificar grupos de diferentes tamanhos e formatos, bem como de densidades diferentes. Como exemplos pode-se citar Markov clustering e Spectral clustering (Zaki and Wagner Meira, 2014).

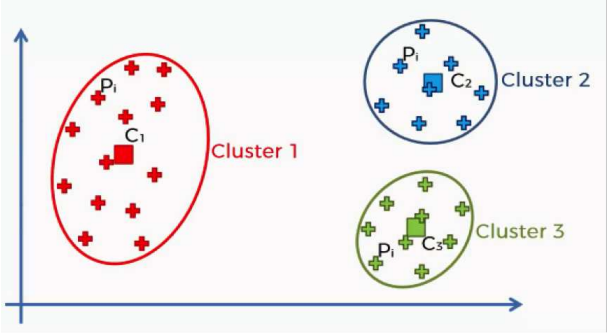


Figura 1.11: Agrupamento pelo método K-means.

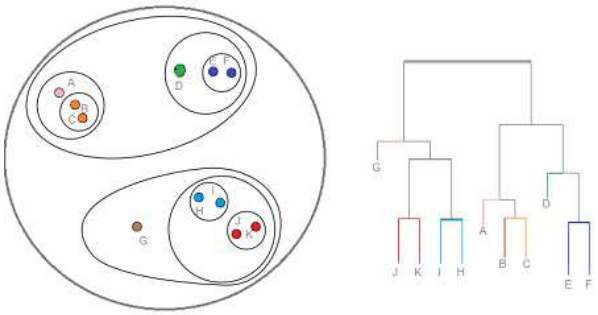


Figura 1.12: Agrupamento pelo método hierárquico.

1.5.1 Agrupamento Espectral

Agrupamento espectral ou spectral clustering, refere-se não somente a um único método, mas a uma família de algoritmos. Em comparação com métodos tradicionais, estes geralmente apresentam melhor performance, e são de fácil implementação, sendo resolvidos de forma eficiente através métodos conhecidos de álgebra linear (Von Luxburg, 2007).

Considerando um conjunto de dados $C = \{x_i\}_{i=1}^{n}$, com n pontos em $\mathbb{R}^d$, denota-se A a matriz $n \times n$ simétrica de similaridade entre os pontos, da seguinte forma:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$$

onde $A(i,j) = a_{ij}$ é a similaridade entre os elementos x_i e x_j, $a_{ij} \geq 0$. A matriz A pode ser considerada também como matriz de adjacência de um grafo ponderado $G(V,E)$, onde cada vértice representa um ponto, e cada aresta conectando os vértices é associado a similidade entre os respectivos vértices.

Estratégias diferentes existem para construir a matriz de similidade. Por exemplo, pode-se conectar todos os pontos do conjunto entre si, produzindo um grafo completamente conectado. Alternativamente, pode-se definir um valor ε, e conectar apenas os vértices que apresentam similaridade menor que esse limite. Ainda, pode-se conectar cada vértice aos k vértices mais próximos. Esta última abordagem é chamada *k-vizinhos mais próximos* (Von Luxburg, 2007).

A base dos algoritmos espectrais são as matrizes Laplacianas dos grafos de similaridade. Na sua forma mais simples, ela é definida como:

$$L = D - A$$

onde D é a matriz diagonal dos graus dos vértices de A,

$$D = \begin{pmatrix} d_{11} & 0 & \dots & 0 \\ 0 & d_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & d_{nn} \end{pmatrix}, d_{ij} = \sum_{j=1}^{n} a_{ij}$$

Outras versões dessa matriz, referidas como matrizes laplacianas normalizadas são:

$$L^s := D^{-1/2} L D^{-1/2}$$

$$L^a := D^{-1} L$$

As matrizes descritas possuem algumas propriedades em comum. Elas são matrizes positivas semi-definidas e possuem n autovalores reais não-negativos: $0 = \lambda_1 \leq ... \leq \lambda_n$. A partir disso, tem-se que a multiplicidade r do autovalor 0 da Laplaciana associada ao grafo G, representa o número de componente conectadas do grafo.

Considere um particionamento $\Gamma = \{C_1, C_2, ..., C_k\}$ do grafo G, tal que $C_i \neq \emptyset$ para todo i, $C_i \cap C_j = \emptyset$ para todo i e j, e $V = \cup_i C_i$. É possível escrever o problema de agrupamento como uma função de minimização das seguintes maneiras:

$$\min_{\Gamma} J_{rc}(\Gamma) = \sum_{i=1}^{k} \frac{W(C_i, \overline{C_i})}{|C_i|} \tag{1.1}$$

$$\min_{\Gamma} J_{nc}(\Gamma) = \sum_{i=1}^{k} \frac{W(C_i, \overline{C_i})}{vol(C_i)} \tag{1.2}$$

onde $W(C_i, \overline{C_i})$ é a soma do peso das arestas entre os vértices de C_i e $\overline{C_i}$, $vol(C_i)$ é a soma dos pesos das arestas com uma extremidade em C_i, e $|C_i|$ é o numero de vértices do conjunto C_i.

Ambas as funções possuem o mesmo principio geral: particionar o conjunto de vértices em k grupos de maneira que as similaridades (aresta) entre vértices de grupos diferentes sejam o menor possível, o que é expresso pela de minimização dos termos $W(C_i, \overline{C_i})$. Em particular, a função 1.1 leva em consideração o tamanho dos grupos ($|C_i|$) como parte do problema de minimização, e tem maior contribuição à medida cresce o número de vértices nos grupos. De maneira similar, a função 1.2 considera a composição do grupo, mas de acordo com o conjunto dos peso das arestas que tem origem no grupo.

Considere $c_i \in \{0,1\}^n$ o vetor que indica a relação de pertinência dos pontos ao grupo C_i, definido da seguinte maneira:

$$c_{ij} = \begin{cases} 1, & \text{se } v_j \in C_i \\ 0, & \text{se } v_j \notin C_i \end{cases} \tag{1.3}$$

Combinando as funções 1.1 e 1.2 com a expressão 1.3, obtém-se as seguintes expressões:

$$\min_{\Gamma} J_{rc}(\Gamma) = \sum_{i=1}^{k} u_i^T L u_i \tag{1.4}$$

$$\min_{\Gamma} J_{nc}(\Gamma) = \sum_{i=1}^{k} u_i^T L^s u_i \tag{1.5}$$

onde $u_i = \frac{c_i}{\|c_i\|}$.

O problema de minimização expresso pelas funções 1.4 e 1.5 é da classe de NP-completo, e portanto, de difícil solução. Isso pode ser contornado relaxando-se as restrições sobre c_i de modo que $c_i \in \mathbb{R}^n$. Através disso, obtém-se como solução para o problema os k menores autovalores da laplaciana utilizada ($0 \leq \lambda_1 \leq ... \leq \lambda_k$) e os seus respectivos autovetores associados ($u_1, ..., u_k$).

No entanto, diferente de c_i, os autovetores gerados não contém informação sobre a relação de pertinência dos pontos aos seus respectivos grupos. Para contornar isso, a matriz $U \in \mathbb{R}^{n \times k}$ dos autovetores é utilizada como uma nova matriz de dados com n pontos num espaço de dimensão k. As linhas dessa matriz são normalizadas para produzir a matriz $Y_{n,k}$, sobre a qual se aplica-se um algoritmo de clusterização, como *K-means*, apenas para designar cada ponto ao seus grupos, que já estão bem separados no auto-espaço gerado Zaki and Wagner Meira (2014). O algoritmo 1, mostra de forma geral o método de agrupamento espectral.

$$U_{n,k} = \begin{pmatrix} | & | & & | \\ u_1 & u_2 & \cdots & u_k \\ | & | & & | \end{pmatrix} \rightarrow Y_{n,k} = \begin{pmatrix} — & y_1 & — \\ — & y_2 & — \\ & \vdots & \\ — & y_n & — \end{pmatrix} \tag{1.6}$$

Algoritmo 1: Agrupamento Espectral

Entrada: $C_{n\times d}$: matriz de dados, k: número de grupos

1 Computar matriz de similaridade $A_{n,n}$
2 Calcular Laplaciana L, L^s, ou L^a
3 Obter os autovetores $\{u_1, u_2, ..., u_k\}$ associados aos k menores autovalores $\{\lambda_1, \lambda_2, ..., \lambda_k\}$ de L
4 $U \leftarrow (u_1, u_2, ..., u_k)$
5 Obter $Y \leftarrow U$ normalizada
6 Computar agrupamento $\Gamma = \{C_1, C_2, ..., C_k\}$ a partir de Y.

1.6 Trabalhos relacionados

Vários trabalhos tem sido propostos no domínio das interações envolvendo proteínas. As abordagens são diversas, usando diferentes atributos em relação ao conjunto dos dados.

O trabalho desenvolvido por Khashan et al. (2012) propôs *SPIDER*, que consiste em uma função de pontuação para experimentos de *docking* baseada em padrões frequentes de interação encontrados em interfaces proteína-proteína. O método utiliza grafos para representar essas interações, que são prospectadas através de análise de contatos no nível de resíduo baseada em triangulação *Almost-Delaunay*. Então, sobre o conjunto de grafos produzidos é aplicado técnicas de mineração para extrair padrões frequentes de interação na interface.

Morozova et al. (2006) usou uma abordagem também baseada em padrões, mas voltados para o domínio das interações RNA-proteína. Nesse caso, dados estruturais foram utilizados a partir de complexos RNA-proteína para extrair padrões em sítios de ligação de nucleosídeos nas proteínas. Um estratégia de superposição foi desenvolvida baseada em interações interatômicas não-covalentes entre cada base do RNA e estruturas de proteínas oriundas do PDB. As interações foram computadas considerando as propriedades físico-químicas dos átomos e a distância entre eles. O método foi capaz de identificar interações frequentes na estrutura dos nucleosídeos, discriminantes em relação a interação de cada base.

O trabalho proposto por Melo et al. (2007), por sua vez, propôs um método baseado em mapas de contatos para obter padrões relevantes em interações proteína-proteína. Os mapas de contatos foram obtidos entre cadeias diferentes, levando em conta as interações interatômicas entre resíduos, e usou técnicas de processamento de imagem para identificar interações conservadas nas interfaces proteína-proteína. O método foi capaz de identificar contatos importantes em complexos para algumas famílias de proteínas.

Finalmente, ppiGReMLIN é uma estratégia baseada em mineração de grafos para inferir padrões de interação na interface de ligação entre proteínas. Ela emprega técnicas de aprendizado de máquina (agrupamento e MSF) sobre um conjunto de grafos (que representam interações em interfaces de PPI's) com o intuito de evidenciar padrões de interação no nível atômico. A estratégia é derivada do método GReMLIN (***GR**aph **M**ining strategy to infer proten-**L**igand **IN**teraction patterns*) Santana et al. (2016), que consiste em uma estratégia para inferir padrões de interação entre proteínas e ligantes (não proteicos). Além disso, ppiGReMLIN é inspirado em trabalhos relacionados a defesa de plantas contra pragas agrícolas. Insetos como a lagarta de *Anticarsia gemmatalis Hübner*, ao atacar as plantas, mais especificamente a soja, ativam a produção de inibidores de proteases por parte da planta, como é o caso do inibidor de proteases do tipo Kunitz (KTI), dificultando o processo de degradação efetuado pelas proteases da lagarta, afetando assim o seu desenvolvimento (Pilon et al., 2017; Patarroyo-Vargas et al., 2017). Motivados pela ação desse inibidor natural, é interesse do grupo de pesquisa propor peptídeos ou peptídeo-miméticos que potencializem esse mecanismo de defesa, afim de promover o controle ecológico mais eficaz deste inseto e outras possíveis pragas na agricultura.

ppiGReMLIN tem sua contribuição como uma técnica que combina modelagem e algoritmos baseados em grafos para prospecção de padrões de interação conservados em interfaces entre proteínas. Além disso, o método usa abordagem em nível atômico, que permite uma descrição mais detalhada das interações, e não depende de alinhamento de sequencia ou superposição de estruturas, podendo ser aplicado para conjuntos de dados com grande volume.

Capítulo 2

Métodos

Nesta seção são descritos os principais processos para a execução da estratégia desenvolvida neste trabalho. Os principais tópicos abordados são o processo de construção das base dados, bem como sua relevância; a modelagem do problema; os processos de mineração de dados; e avaliação geral da estratégia.

O fluxograma geral da estratégia é mostrado na figura 2.1, e pode ser descrito sumariamente em três etapas:

(A) **Aquisição de dados e modelagem**: extração de dados de complexos proteína-proteína a partir do PDB, prospecção de interações nas interfaces desses complexos e sua representação na forma de grafos;

(B) **Agrupamento**: representação da base de dados de grafos na forma de matriz de dados, redução de dimensionalidade através de descomposição em valores singulares (SVD), agrupamento dos dados e sua avaliação;

(C) **Mineração de subestruturas frequentes**: mineração de subestruturas frequentes nos grupos resultantes da etapa anterior e mapeamento dos padrões para o conjunto de grafos.

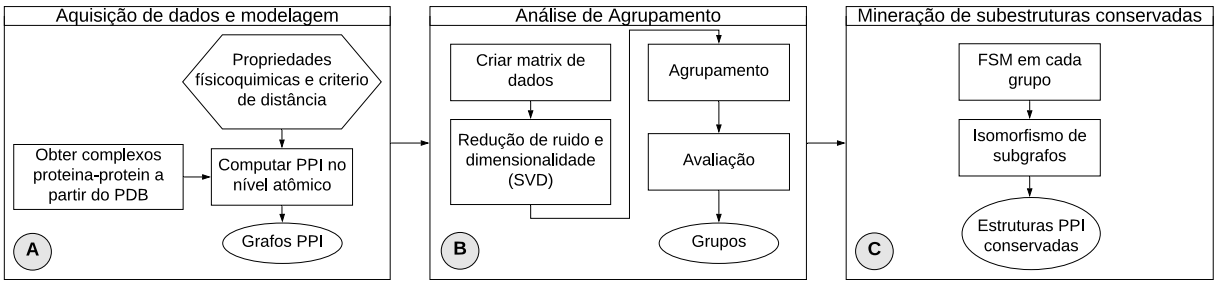


Figura 2.1: Fluxograma da estratégia ppiGReMLIN

2.1 Aquisição de dados e modelagem

O primeiro passo da estratégia consiste na obtenção dos dados a partir do *Protein Data Bank* (PDB). Esta é uma base de dados biológica composta de estruturas tridimensio-

nais de proteínas e outras moléculas complexas, como ácidos nucleicos. As estruturas depositadas no PDB são obtidas experimentalmente através de técnicas como cristalografia de raios X e ressonância magnética nuclear, e descrevem as estruturas de proteínas segundo a sequência de seus aminoácidos, estrutura secundária, composição atômica, etc. No nível atômico é possível obter informações diversas, dentre as quais destaca-se o posicionamento dos átomos na proteína em coordenadas tridimensionais (x, y, z) no espaço. Cada estrutura resolvida no PDB possui um identificador único, o PDBid, e pode conter uma ou mais proteínas além de outras moléculas.

No escopo da proposta do ppiGReMLIN, as interações são computadas em nível atômico nas interfaces de complexos proteína-proteína. Em geral, a cada componente de um complexo obtido do PDB atribui-se um identificador de cadeia diferente, representado por uma letra. Nesse sentido, define-se como interface a região de contato entre os componentes representados por cadeias diferentes, e assim as interações são computadas apenas entre pares de átomos de cadeias distintas. A estrutura representada na figura 2.2 mostra a interface entre os componentes da estrutura de PDBid 3D65. As cadeias E e I correspondem cada, a um componente diferente do complexo, e são referidas respectivamente com o seu PDBid como 3D65:E e 3D65:I.

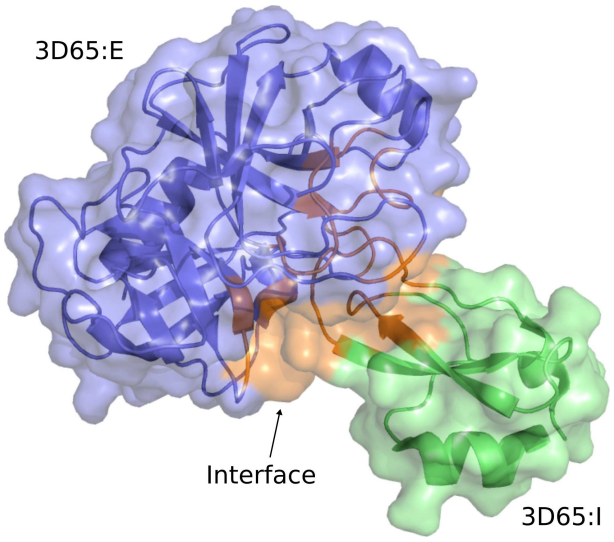


Figura 2.2: Interface em complexos proteína-proteína

2.1.1 Grafos de contatos

Contatos, em bioinformática, refere-se a uma abordagem geométrica para definição de interações entre entidades (átomos, resíduos, e outros). Essencialmente, os métodos são classificados como dependentes ou livres de *cut-off* (da Silveira et al., 2009). No primeiro caso, a determinação do contato é fundamentado em torno de um critério limítrofe de distância entre as entidades, enquanto que no último, esse critério não é necessário.

Neste trabalho, utiliza-se uma abordagem dependente de *cut-off* da seguinte maneira: dois átomos i e j estão em contato se j está inserido dentro de uma esfera centrada em i com raio r, onde r é chamado distância de *cut-off*. Nesse caso, i e j são átomos de cadeias distintas, como descrito anteriormente na definição de interface.

Uma vez determinados, os contatos são usados como base para construção de um conjunto de grafos representando as interações nas interfaces dos complexos proteína-proteína. Nesse sentido, considerando um grafo $G(V, E)$, cada vértice de V representa um átomo, e as arestas entre os vértices representam as interações entre os átomos. Cada vértice é rotulado de acordo com as propriedades físico-químicas dos seus respectivos átomos, aos quais é atribuído uma combinação dos seguintes rótulos: aceptor (ACP), aromático (ARM), doador (DON), hidrofóbico (HPB), negativo (NEG) e positivo (POS) [1]. Em seguida, as interações (arestas) são computadas de acordo com um critério de distância e com base nos rótulos dos respectivos átomos (tabela 2.1), e assim, cada aresta do grafo recebe, igualmente, uma combinação de rótulos. Estes são: empilhamento aromatico, ligação de hidrogênio, hidrofóbica, repulsiva e ponte salina (Santana et al., 2016; Gonçalves-Almeida et al., 2011; Silveira et al., 2014; Fassio et al., 2017, 2018).

Tabela 2.1: Tipos dos átomos e critério de distância (em Å) para computar interações.

Tipo da interação	**Tipos dos átomos**	**Distância**	
		MIN	MAX
Empilhamento Aromático	2 átomos aromáticos	1.5	3.5
Ligação de Hidrogênio	1 átomo aceptor e 1 átomo doador	2.0	3.0
Hidrofóbica	2 átomos hidrofóbicos	2.0	3.8
Repulsiva	2 átomos com mesma carga	2.0	6.0
Ponte Salina	2 átomos com cargas opostas	2.0	6.0

A figura 2.3 mostra um exemplo de grafo gerado na estrutura com PDBid 1SLW. Com relação à designação de rótulos, é importante ressaltar que um rótulo é atribuído a um vértice somente quando este contribui para composição de alguma interação. Por exemplo, o vértice GLU39:OE1 na figura 2.3, tem como possíveis rótulos os atributos *aceptor* e *negativo*. No entanto, na constituição da interação do tipo *ponte salina* com o vértice LYS97:NZ, apenas o atributo *negativo* oferece alguma contribuição, de maneira que o vértice recebe como rótulo apenas este atributo.

Outro ponto importante diz respeito ao número de grafos produzidos em cada complexo proteína-proteína. As interações podem ocorrer em toda extensão na in-

[1] É importante ressaltar que as palavras "*rótulo*" e "*tipo*" são usadas como sinônimos quando se referem aos atributos dos vértices (átomos) e das arestas (interações) nos grafos de contatos. Assim, um átomo (ou vértice) do tipo X, especifica a designação do atributo X como um rótulo do respectivo átomo ou vértice. O mesmo se aplica às arestas (ou interações)

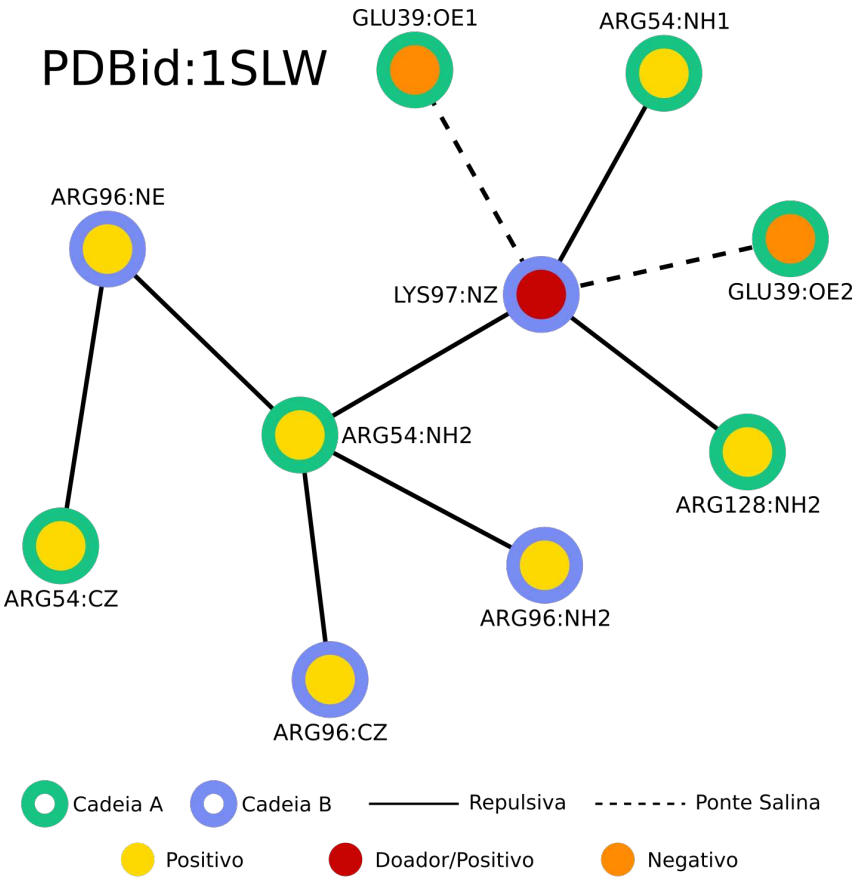


Figura 2.3: Conjunto de grafos de contato

terface de contato, muitas vezes gerando grafos desconexos. Em vista disso, para a estratégia geral, considera-se como grafo cada uma dessas componentes separadamente.

Como consequência direta da maneira como as interações são computadas, apenas entre átomos de cadeias diferentes, segue que os grafos gerados são bipartidos. Um grafo bipartido é da forma $G(P, I, E)$, cujo conjunto de vértices pode ser particionado em dois conjuntos disjuntos P e I, e cujas arestas em E possuem uma das extremidades em P e a outra I. Isso pode ser observado no grafo da figura 2.3, onde dois vértices com bordas de cores iguais, jamais se conectam.

2.2 Agrupamento

O intento desta etapa é tomar o conjunto de grafos oriundos da prospecção de contatos e organizá-los em grupos similares segundo as propriedades físico-químicas de seus vértices e arestas, e sua topologia.

2.2.1 Matriz de dados

Afim de agrupar o conjunto de grafos, é necessário primeiro representa-los na forma de uma matriz de dados. Para isso foi proposta uma matriz de contagem onde cada grafo é representado segundo rótulos de cada um de seus pares de vértices adjacentes.

Considere um grafo G_1 (figura 2.4) com vértices A, B e C, rotulados *doador* (DN), *aceptor/negativo* (AC/NG) e *doador/positivo* (DN/PS), respectivamente, e arestas AB e AC, representando uma ligação de hidrogênio e uma ponte salina, respectivamente. Para representar G_1, cada um de seus pares de átomos adjacentes é rotulado através da união dos rótulos em seus vértices. Assim, os pares de átomos AB e BC, por exemplo, são rotulados DN-DN/PS e AC/NG-DN/PS respectivamente. Em seguida, cada rótulo distinto dá origem a uma coluna na matriz de contagem, e para cada grafo ou linha desse matriz, é registrado o número de ocorrências dos pares com esses rótulos em sua estrutura. A tabela 2.2 ilustra a matriz de contagem gerada a partir dos grafos da figura 2.4.

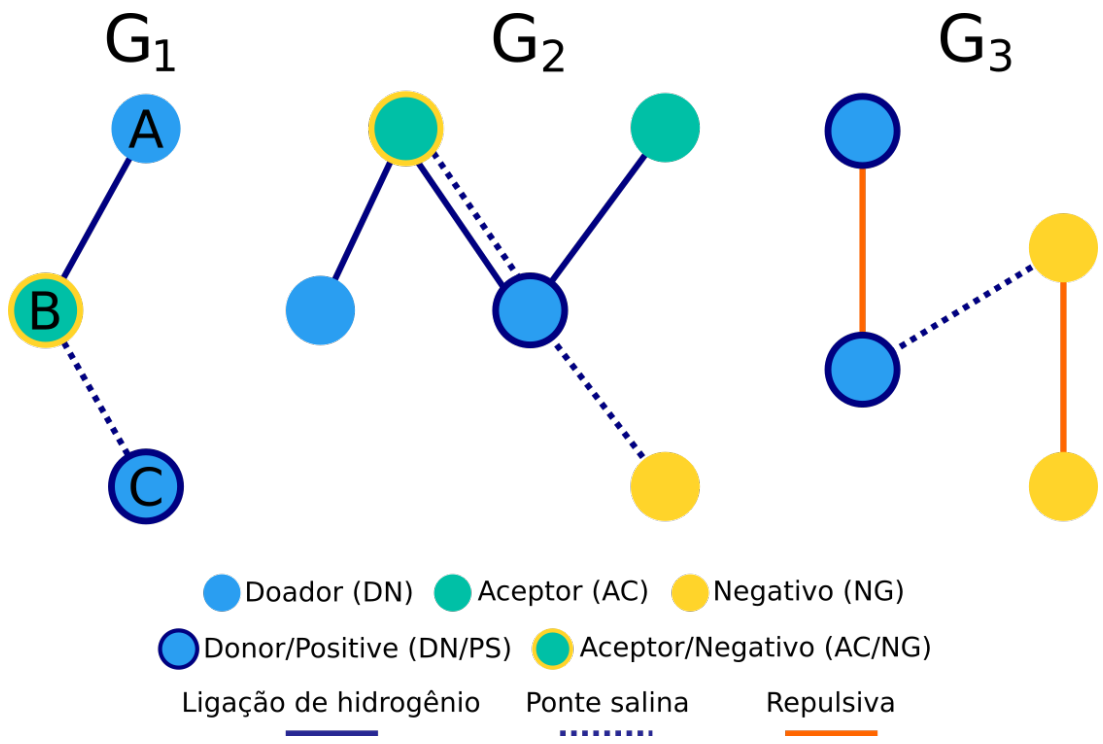


Figura 2.4: Exemplo de grafo de contato

2.2.2 Redução de dimensionalidade e ruído

Nessa etapa, a matriz de dados é simplificada através de decomposição em valores singulares (SVD). Isso resulta em uma representação mais compacta dos dados de entrada, que pode ser manipulada de maneira mais eficiente pelos métodos de agrupamento, em relação ao uso de memória e ao tempo de execução (Eldén, 2006).

Tabela 2.2: Exemplo de matrix de contagem.

Grafo	AC-DN/PS	AC/NG-DN	AC/NG-DN/PS	DN/PS-NG	DN/PS-DN/PS	NG-NG
G_1	0	1	1	0	0	0
G_2	1	1	1	1	0	0
G_3	0	0	0	1	1	1

Seja uma matriz $X_{l\times c}$, com posto r. SVD decompõe X no produto $U\Sigma V^T$, onde: $U_{l\times r}$ é uma matriz ortogonal, cujas colunas são os vetores singulares à esquerda de X; $\Sigma_{r\times r}$ é uma matriz diagonal, cujos elementos da diagonal, chamados valores singulares, são positivos e estão ordenados em ordem decrescente; e $V_{c\times r}$ é uma matriz ortogonal, cujas colunas são os vetores singulares à direita de X.

Para reduzir o nível de ruído, primeiro a matriz de dados X é aproximada pela da matriz X_d, com posto d ($d < r$), obtida pelo produto $U_d\Sigma_d V_d^T$, onde U_d, Σ_d e V_d, são as matrizes formadas pelas primeiras d colunas de U, Σ e V, respectivamente. É possível mostrar que $X_d = U_d\Sigma_d V_d^T$ é a melhor aproximação de X, com posto d, no que diz respeito ao erro quadrático médio (Leskovec et al., 2014). Em seguida, para a redução de dimensionalidade, toma-se o produto $U_d\Sigma_d$ a partir de X_d, que resulta em uma matriz de dimensão $m \times d$, que é uma aproximação de X com menos colunas. A idéia por trás disso vem do fato de que as colunas de U_d capturam padrões entre os pontos em X_d (Tan, 2006).

A determinação do valor adequado do parâmetro d foi feita mediante a aṅalise da distribuição dos valores singulares. Para isso foi considerado um valor mínimo de 95% da variância em relação a matriz original de dados.

2.2.3 Agrupamento e estratégia de avaliação

O agrupamento dos grafos foi realizado através do algoritmo de agrupamento espectral. Para isso, foi utilizada a implementação disponível no pacote *scikit-learn* (versão 0.19.2) em *Python*.

A matriz de similaridade foi construída com a abordagem *k-vizinhos mais próximo* (*k-nearest neighbors*), sendo este, um parâmetro do entrada do método, junto com a métrica empregada no calculo da similaridade entre pontos da matriz de dados. Como métrica, foi escolhida a norma L_2, ou simplesmente, distância euclidiana entre os pontos.

Os últimos parâmetros a serem definidos na estrategia foram o número de grupos n e o número de vizinhos k da abordagem *k-vizinhos mais próximos*. Dependendo do valor de k, o grafo de similaridade produzido pode ser conexo ou não. No entanto, o método espectral considera cada componente desconexa como um grupo separado, o que é adequado somente quando há algum conhecimento prévio a respeito do conjunto de dados que indique tal agrupamento. Em contrapartida, quando esse não é o caso, tem-se como bom principio escolher um valor de k tal que o grafo de similaridade seja conexo (Von Luxburg, 2007). Nesse sentido, k foi determinado experimentalmente, variando o seu valor no intervalo $\{\lfloor 0.01r \rfloor, \lfloor 0.02r \rfloor, ..., \lfloor 0.9r \rfloor\}$, onde r é do numero de pontos da respectiva matriz de dados. O valor final foi escolhido como o menor valor dentro do intervalo, para o qual o grafo é conectado.

Com relação ao número de grupos n, este foi determinado de acordo com a heurística *eigen-gap*, pela maximação da diferença de autovalores consecutivos das respectivas matrizes laplacianas associadas.

2.3 Mineração

Nesse ponto, em cada grupo gerado objetivou-se encontrar estruturas frequentes, que foram posteriormente mapeados para o conjunto de grafos de contatos prospectados na primeira etapa do ppiGReMLIN.

2.3.1 Mineração de subgrafos

Para a mineração de subestruturas conservadas, foi utilizado o algorítimo *gSpan* (Yan and Han, 2002). O método toma como parâmetros de entrada o conjunto de grafos e o suporte mínimo (s_{min}) para encontrar padrões.

Nesta etapa, o conjunto de grafos foi convertido em multigrafos. Para isso, as arestas com múltiplos rótulos foram desdobradas em múltiplas arestas, cada uma contendo um rótulo. O objetivo disso é capturar um maior número de padrões nas estruturas dos grafos, uma vez que o *gSpan* não consegue distinguir atributos individuais em arestas com múltiplos rótulos.

A mineração de padrões foi feita com valores de suporte no intervalo $\{0.1, 0.2, ..., 1.0\}$. Um dos problemas do *gSpan*, refere-se ao fato de ele produzir um número muito grande de padrões, dos quais grande parte estão contidos em padrões maiores. Nesse sentido, em cada grupo, para cada suporte, os grafos maximais foram computados, e apenas estes foram considerados como padrões finais. Além disso, para efeito de análise, apenas os dez maiores padrões maximais foram considerados como padrões maximais finais.

Finalmente, um valor de suporte foi escolhido com base na análise do tamanho

dos padrões maximais, assim como no número de padrões gerados. Essa escolha leva em conta o fato de que padrões maiores ocorrem com menor frequência, e buscou-se balancear essas duas tenências opostas.

2.3.2 Isomorfismo de subgrafos

Isomorfismo de subgrafos é central para duas tarefas: computar os grafos maximais e mapear os padrões para seus respectivos grafos de origem.

Nesse ponto, foi utilizado o algoritmo *VF2* (Cordella et al., 2004). No entanto, implementações disponíveis deste algoritmo executam a tarefa de isomorfismo de subgrafos apenas para subgrafos induzidos por vértices, e os grafos gerados pela mineração de subgrafos frequentes nem sempre atendem a este requisito, contendo apenas um subconjunto das arestas dos seus supergrafos entre pares de vértices correspondentes nos respectivos padrões.

Para contornar esse problema, os grafos gerados (padrões e grafos de contatos) foram primeiro transformados em grafos linha, e então foi executada a tarefa de isomorfismo. A base para isso vem das seguintes proposições (Harary (1969)):

- dois grafos H e J isomórficos, apresentam grafos linha isomórficos;
- se os grafos linha de dois grafos H e J são isomórficos, então H e J também são, com apenas algumas exceções (que não se aplicam ao escopo deste trabalho, onde os grafos são todos bipartidos).

.

Além disso, os padrões, que são subgrafos induzidos por arestas dos grafos de contatos, tem seus grafos linha como subgrafos induzidos por vértices em relação aos grafos linha dos seus respectivos supergrafos, e assim, os grafos linha podem ser utilizados diretamente como entrada para a tarefa do isomorfismo de subgrafos.

A figura 2.5 ilustra a relação entre subgrafos induzidos e o seus respectivos grafos linha, e o isomorfismo de subgrafos. Observa-se, por exemplo, que G' e G'' são subgrafos do grafo G, mas não são induzidos por vértices, pois não tem as arestas 1 e 5, respectivamente. No entanto, os seus respectivos grafos linha são induzidos por vértices em relação ao grafo linha $L(G)$.

É importante observar ainda que a definição de grafos linha utilizada neste trabalho aplica-se somente a grafos simples (sem arestas múltiplas ou laços), e não é válida para os grafos de contatos ou padrões gerados, que são multigrafos. Assim, antes de serem transformados em grafos linha, eles são remodelados na forma de grafos simples, e as arestas múltiplas são representadas como arestas simples, onde os seus respectivos atributos são condensados em um único atributo. A partir daí, o isomorfismo de subgrafos é executado, e cada atributo dos rótulos múltiplos, agora sobre os

vértices, podem ser distinguidos pela utilização de regras semânticas na comparação dos rótulos.

A figura 2.6 mostra um exemplo de mapeamento de padrão para um grafo de entrada. Observa-se a possibilidade de um padrão ser mapeado de maneiras diferentes. Observa-se também que vértices e arestas mapeados podem conter apenas um subconjunto dos atributos nos vértices e arestas correspondentes nos grafos de entrada, o que é ilustrado através do mapeamento do vértice *a* (hidrofóbico) para o vértice 2 (aromático/hidrofóbico), e da aresta *a-b* (hidrofóbica) para a aresta *2-3* (aromática/hidrofóbica), no mapeamento 1.

Para a tarefa de isomorfismo de subgrafos, foi utilizado a implementação do algoritmo *VF2* no pacote *NetworkX* (Hagberg et al., 2008) versão 2.3 em Python. Além disso, toda a modelagem de estruturas envolvendo grafos e grafos linha, com exceção da mineração de subgrafos, foi feita utilizando este pacote.

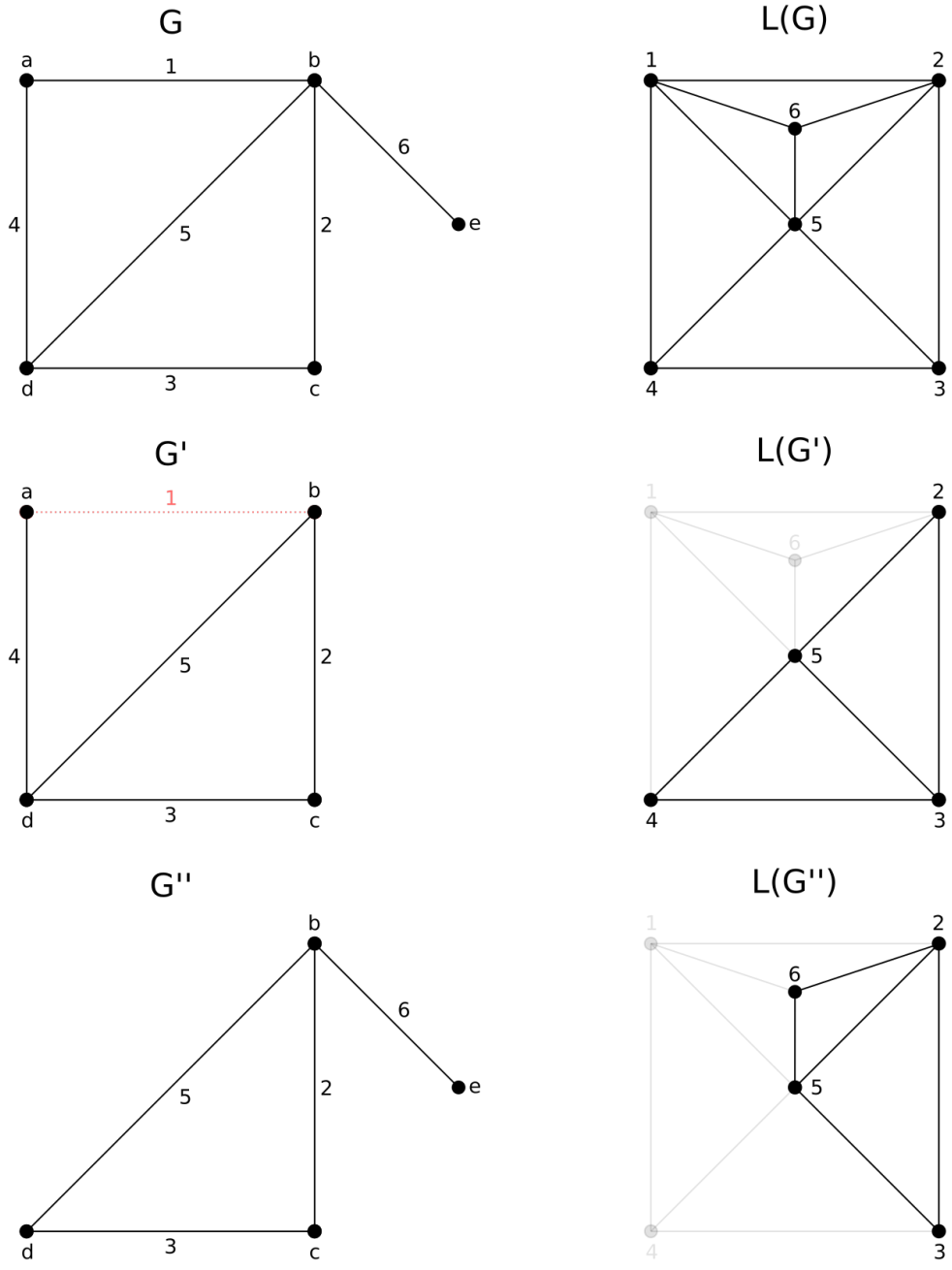


Figura 2.5: Exemplo de isomorfismo de subgrafos e grafos linha

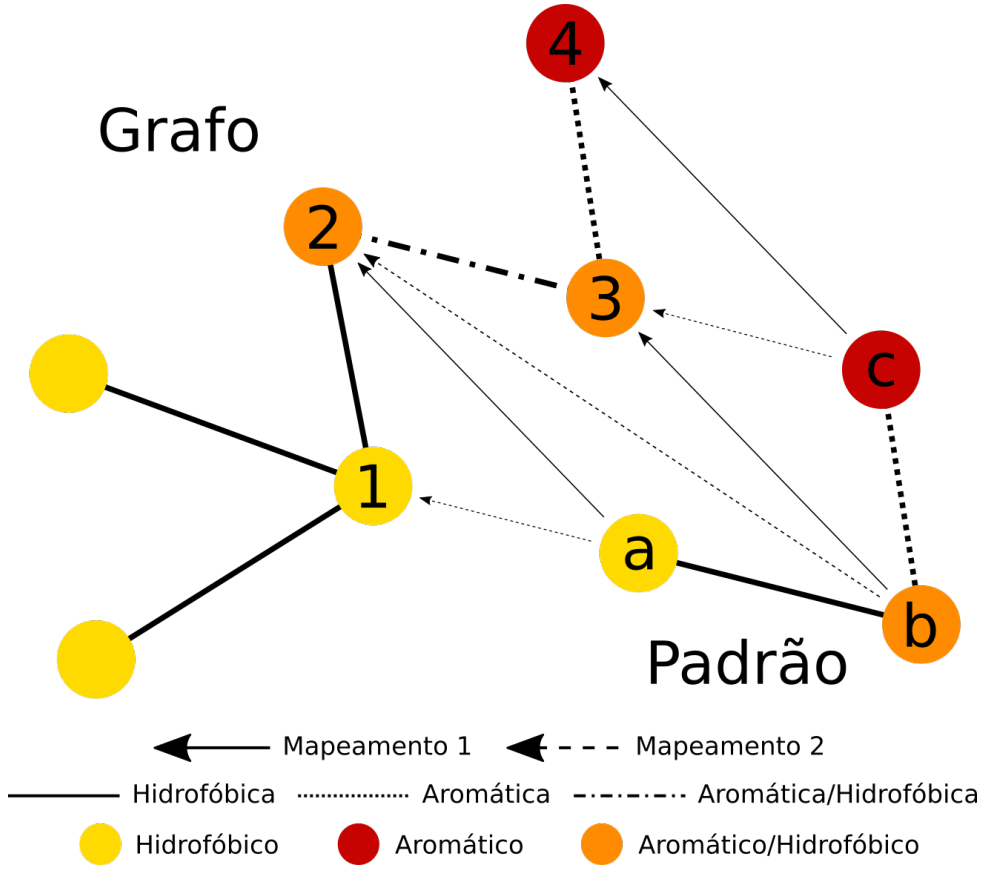


Figura 2.6: Exemplo de mapeamento de um padrão para um grafo entrada

Capítulo 3

Experimentos e Resultados

3.1 Conjuntos de dados

A estratégia proposta foi instanciada com dois conjuntos de dados contendo complexos proteína-proteína obtidos a partir do PDB: Serino Protease (SP) e BCL2. O conjunto SP é constituído por proteases da família das tripsinas ou *tripsinas-like*, associado a um inibidor, e o conjunto BCL2 é constituído por complexos da família de proteínas *BCL-2*. Abaixo é descrita a motivação para a escolha de cada conjunto de dados.

3.1.1 Serino protease

A soja é uma cultura de grande importância econômica para o Brasil, sendo a principal cultura do agronegócio brasileiro. Segundo os dados da safra de 2016/2017, o Brasil sozinho é responsável por um terço de toda a produção mundial do grão, com aproximadamente 113 milhões de toneladas, perdendo apenas para os Estados Unidos com 117 de milhões de toneladas (emb, 2017). No entanto, considerando-se as taxas de crescimento da produção, espera-se que em dez anos, o país se torne o maior produtor mundial. Segundo dados de 2016, a soja ocupou a primeira posição dos produtos exportados pelo país, correspondendo a uma total de aproximadamente U$ 19,3 bilhões o que representa 10% de toda a exportação brasileira (oec, 2017).

No entanto, um dos entraves para o aumento da produção mundial da soja esta relacionado a existência de pragas (como insetos e vários agentes patógenos) que reduzem a produção das safras, mesmo considerando o uso significativo de pesticidas químicos e de tecnologias alternativas como transgenia.

A lagarta da soja, *Anticarsia gemmatalis*, é uma praga desfolheadora, e uma das principais pragas que atacam a soja no Brasil, sendo que mesmo em baixas densidades populacionais pode causar prejuízos à lavoura, que podem levar a sua destruição (Silva et al., 2002).

O combate à esse tipo de praga normalmente é feito utilizando-se agroquímicos. No entanto, tem-se buscado métodos alternativos, livres desses compostos, em razão

do seu alto custo de desenvolvimento para combater pragas cada vez mais resistentes, e pela pressão de questões ambientais e do consumo de alimentos livre de agrotóxicos.

Uma das abordagens para lidar com esse problema consiste em explorar os próprios mecanismos de defesa das plantas contra o ataque de pragas. No caso dos insetos, um dos mecanismos mais diretos de defesa das plantas, envolve a potencialização de inibidores de proteases. As proteases são as enzimas responsáveis pela degradação de proteínas da dieta dos insetos, e a sua inibição prejudica o desenvolvimento do inseto, pois interfere no processo de degradação de proteínas ingeridas, provocando uma deficiência na disponibilidade de aminoácidos para a síntese das proteínas do seu próprio organismo (Scott et al., 2010; Wielkopolan et al., 2015).

No caso da lagarta da soja, assim como em muitas outras espécies de lepidópteras, as serino-proteases são a principal classe de enzimas na região do intestino médio (Terra and Ferreira, 1994). Desse ponto de vista, é fundamental entender os mecanismos de interação entre as serino-proteases e o seus respectivos inibidores, para desenvolvimento de outros inibidores orgânicos, proteicos ou não-proteicos, para serem utilizados no controle de pragas.

O ponto de partida para a construção do conjunto de dados Serino Protease foi a sequencia obtida de uma protease digestiva da lagarta da soja em trabalhos anteriores (Pilon et al., 2013), disponível na base de dados GenBank, código JX898746.1 [2013]. A partir desta sequência, buscou-se no PDB estruturas com similaridade de sequencia mínima de 30% em relação à sequencia modelo. O resultado foi filtrado, retendo-se apenas as estruturas representando complexos proteína-proteína. O conjunto final produzido é composto por 93 estruturas, a maioria contendo alguma serino-protease associada a algum tipo de inibidor.

3.1.2 BCL2

Apoptose é o processo de morte celular programada que ocorre naturalmente nos organismos. Está envolvida no desenvolvimento e envelhecimento dos organismos, na manutenção da homeostase celular, e nos mecanismos de defesa, por exemplo, através da remoção de células danificadas por doenças ou agentes nocivos. Pode ocorrer de forma extrínseca, ativada pela associação de ligantes moleculares à receptores na superfície celular, ou de forma intrínseca, ativada pela liberação do citocromo *c* do interior das mitocôndrias. (Elmore, 2007)

As proteínas da família BCL-2 tem um papel importante na regulação da apoptose. Elas são descritas através da homologia em relação aos quatro domínios presentes na proteína BCL-2 (*B-cell lymphoma protein 2*), o que deu nome ao grupo. As proteínas dessa família são classificadas como anti-apoptóticas ou pro-apoptóticas. As proteínas anti-apoptóticas, em geral, contém os quatro domínios (BH1-4), como *BCL*-2 e

BCL-x$_L$, ou três dominios (BH1-3), como *MCL*-1, e agem associando-se as proteínas pro-apoptóticas, inibindo sua atividade e preservando a integridade estrutural da membrana externa das mitocôndrias. As pro-apoptóticas por sua vez podem conter múltiplos domínios, como BAK, BAX e BOK, ou apenas o domínio BH3, como BIM, PUMA e NOXA. Em última instancia, os estado da célula depende do equilíbrio relativo entre esse dois tipos de proteínas (Bhat et al., 2013; Petros et al., 2004).

Uma variedade de doenças está associada diretamente a distúrbios na regulação da apoptose, que podem ocorrer desde a fase embrionaria até a fase adulta (Sorenson, 2004). Um dos casos com maior evidência está relacionada a diversos tipos de câncer. De fato, apoptose é considerada como um mecanismo fundamental para a supressão da tumorigênese (Delbridge et al., 2012). Por outro lado, a evasão das células aos mecanismos apoptóticos juntamente com mutações oncogênicas que desregulam o crescimento e o ciclo celular, favorecem de maneira profusa esse processo (Delbridge and Strasser, 2015), o que caracteriza a evasão da apoptose entre os *"Hallmarks of Cancer"* (Hanahan and Weinberg, 2011). Nesse sentido, existem vários estudos voltados para o desenvolvimento de miméticos do BH3 e alguns testes clínicos relacionados a via apoptótica BCL-BH3 Delbridge et al. (2016).

Para o conjunto de dados BCL2, buscou-se no PDB por sequências com mínimo de 30% de similaridade em relação a sequencia da proteína *MCL*-1, extraída da estrutura com PDBid 2KBW. A partir disso, os resultados foram filtrados de modo a obter apenas aqueles contendo complexos proteína-proteína, o que produziu um conjunto final com 72 estruturas.

3.2 Grafos e matrizes de dados

Como resultado da prospecção de contatos sobre as estruturas extraídas do PDB, obteve-se um conjunto de 875 grafos para a base de dados *SP*, e 1188 grafos para a base de dados *BCL*2. O número de colunas das respectivas matrizes de dados geradas foram 24 e 21.

A definição do número de colunas d da forma reduzida das matrizes dos conjuntos de grafos foi feita experimentalmente junto com a análise de agrupamento descrito na seção a seguir.

3.3 Análise de agrupamento

Os parâmetros do agrupamento foram definidos com base na heurística *eigen-gap*. Para isso, tomou-se k (número de vizinhos) como o menor valor dentro do intervalo $\{\lfloor 0.01r \rfloor, \lfloor 0.02r \rfloor, ..., \lfloor 0.9r \rfloor\}$ — onde r é o numero de linhas da respectiva matriz —

para o qual o grafo de similaridade é conectado. Isso foi feito para cada forma reduzida da respectiva matriz de dados com número de colunas d, obtendo-se um valor n (numero de grupos) para cada uma destas. Os resultados são mostrados na figura 3.1 para cada conjunto de dados.

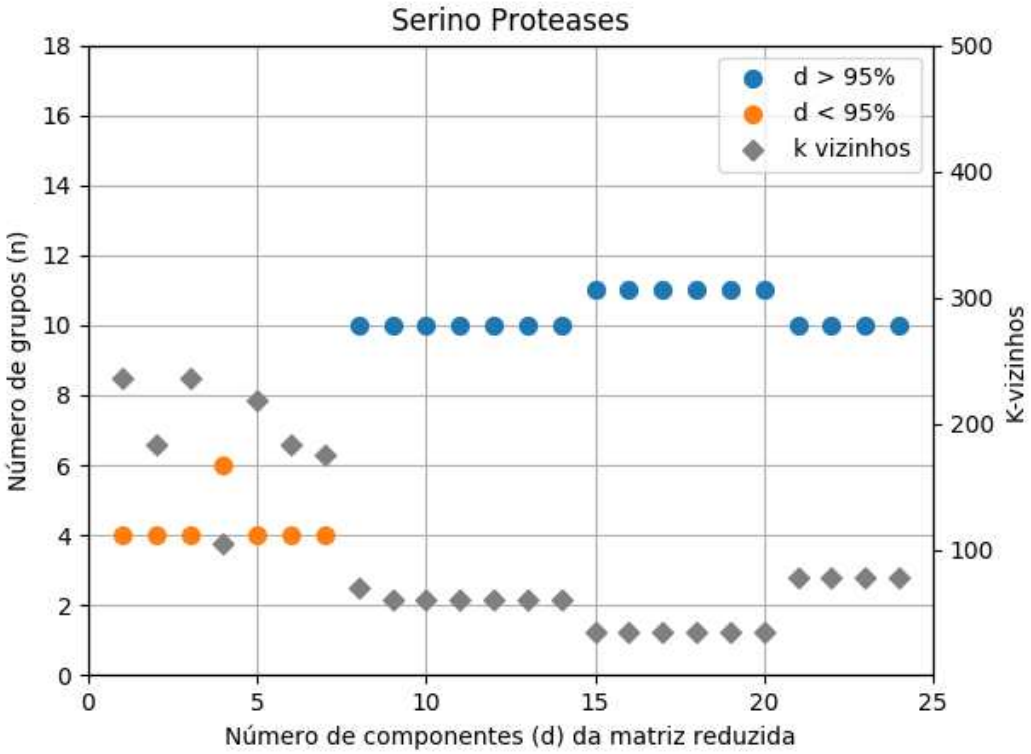
Para a base de dados SP, o menor valor para o número de vizinhos em todo o experimento foi 35, ocorrendo para matrizes com numero de colunas d no intervalo $I_{SP} = [15, 20]$. Observa-se que neste intervalo o número de grupos é constante e igual a 11, e a composição da variância em relação à matriz original é acima de 95%. Nesse caso, o valor de d foi escolhido como $d = 15$, que representa a forma mais simplificada da matriz de dados dentro do intervalo I_{SP}, para a qual os outros parâmetros são constantes. Assim o valor final para os parâmetros é $(n, k, d) = (11, 35, 15)$.

Para a base de dados BCL2, o menor valor para o número de vizinhos em todo o experimento foi 23. No entanto este ocorre para uma matriz, com $d = 3$, cuja composição de variância é abaixo de 95% da matriz original, e portanto foi desconsiderado. No espectro válido das matrizes reduzidas, o menor valor para o parâmetro k é 35, para as quais o valor de d se encontra no intervalo $I_{BCL2} = [10, 24]$. Ainda, sobre a maior parte deste intervalo, o número de grupos permanece constante e igual 13. Nesse caso, o valor de d foi escolhido como $d = 10$, para o qual $n = 13$, que representa a forma mais simplificada da matriz de dados dentro do intervalo I_{BCL2}. Assim o valor final para os parâmetros é $(n, k, d) = (13, 35, 10)$.

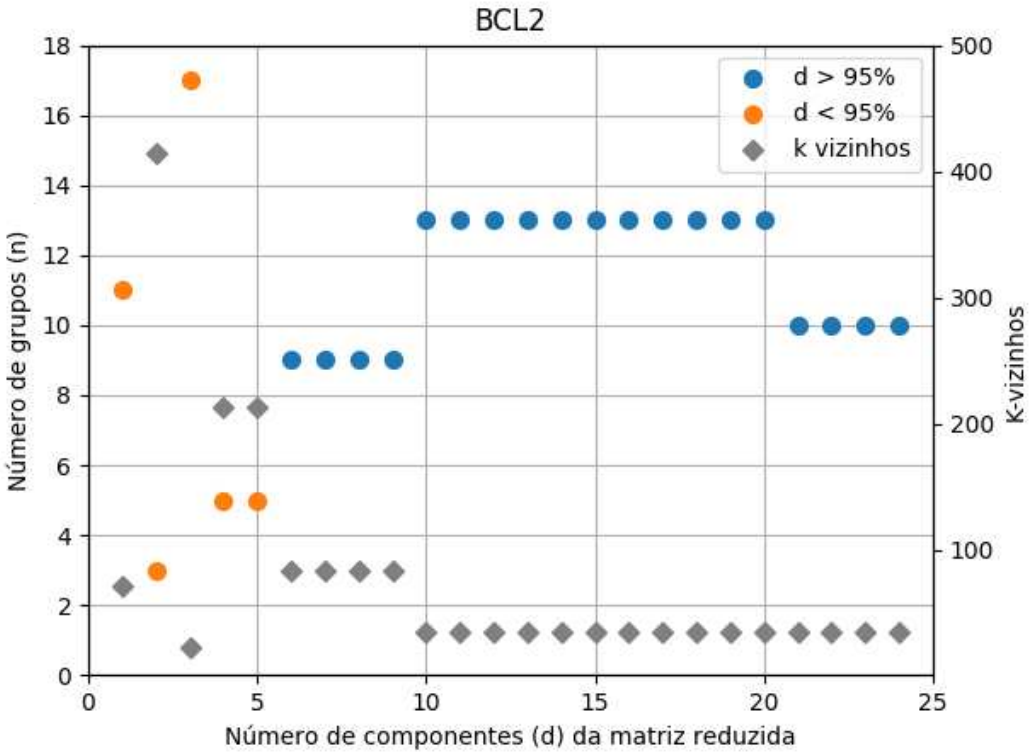
3.4 Mineração de subgrafos

A escolha de um valor de suporte apropriado para a definição de padrões relevantes foi feita com base na análise de tamanho dos padrões maximais produzidos e na sua ocorrência nos grupos. Como principio geral foi definido um valor de suporte mínimo igual a 0.5. A partir disso, foi criada uma tabela que mostra os atributos para cada suporte em cada grupo. A figura 3.2 mostra um trecho da tabela onde são analisados os padrões para cada conjunto de dados. O maiores padrões (maior número de vértices) foram encontrados nos suportes 0.6 e 0.7, e por isso apenas eles são ilustrados.

Para a base de dados SP, o suporte escolhido foi 0.7. Nesse caso, os maiores padrões do conjunto estão no grupo 6, e no suporte em questão, estes apresentam três padrões diferentes, cada um com seis vértices. Além disso, dois dentre estes padrões apresentam suporte real igual 0.83 (33 de 40 estruturas). No suporte 0.6, o número de padrões do grupo reduz para dois, cada um com sete vértices. Considerando a diferença de apenas um vértice entre os suportes, e o maior número de padrões, levando em conta ainda a menor ocorrência no suporte menor, achou-se apropriada a escolha para o suporte mínimo o valor 0.7.



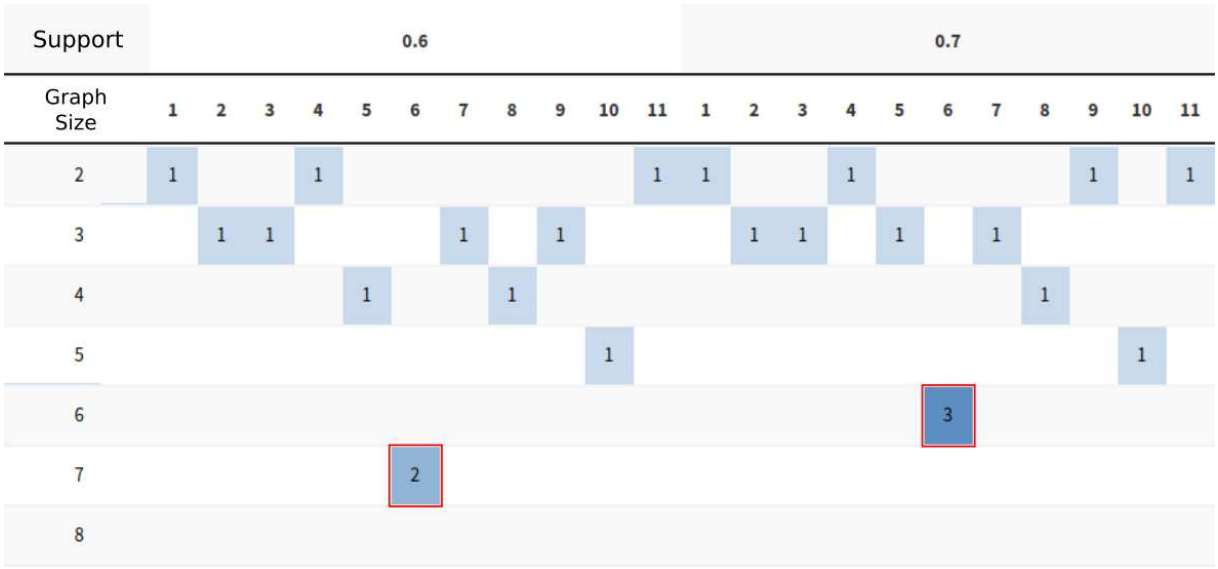
(a)



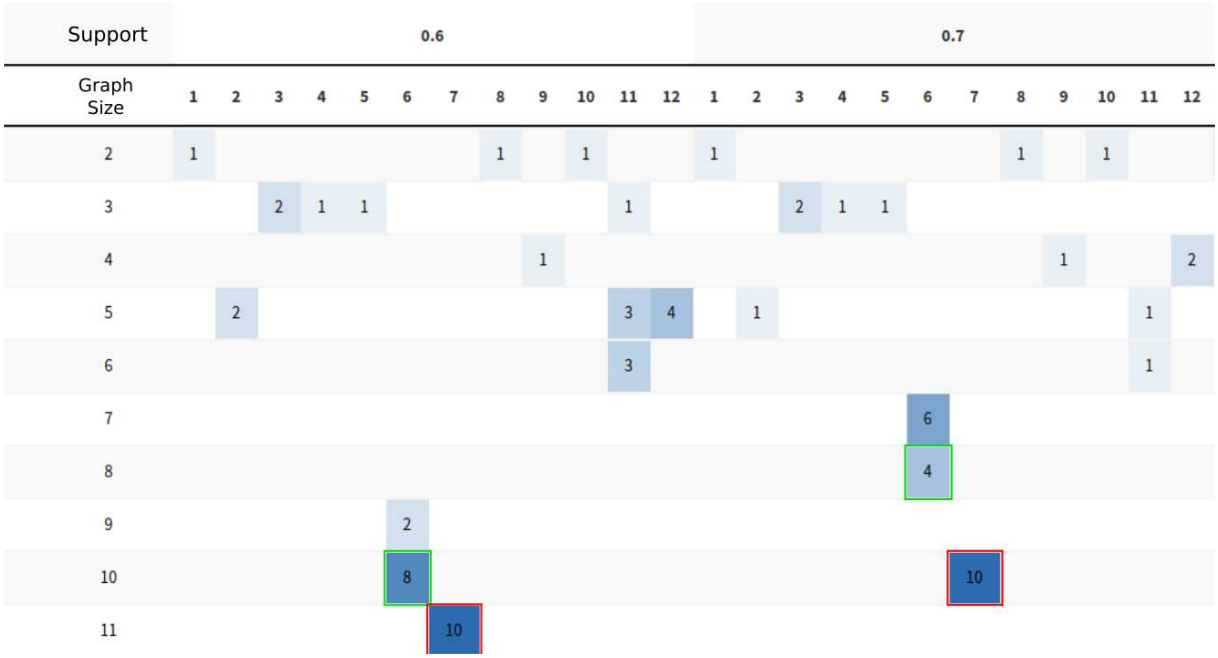
(b)

Figura 3.1: Análise de agrupamento para os conjuntos de dados Serino Proteases e BCL2

Para a base de dados BCL2, o suporte escolhido foi 0.6. Nesse caso, os grupos mais expressivos em termos do tamanho dos seus respectivos padrões foram os grupos 6 e 7. Por este critério, no suporte 0.7, os padrões somam um total de dez estruturas, com onze e dez vértices respectivamente em cada grupo, que supera o tamanho dos padrões em relação ao suporte 0.6 por dois vértices e um vértice, no mínimo, respectivamente nos mesmos grupos. Por essa razão, o suporte escolhido para a base de dados BCL2 foi 0.6



(a) Serino Proteases



(b) BCL2

Figura 3.2: Análise de padrões para os conjuntos de dados Serino Proteases e BCL2 em diferentes suportes

3.5 Estruturas Conservadas

Neste seção são discutidas as características principais dos padrões considerados mais interessantes para cada base de dados nos seus respectivos suportes escolhidos.

3.5.1 Serino Proteases

A figura 3.3 mostra exemplos de padrões e alguns grafos onde estes são encontrados.

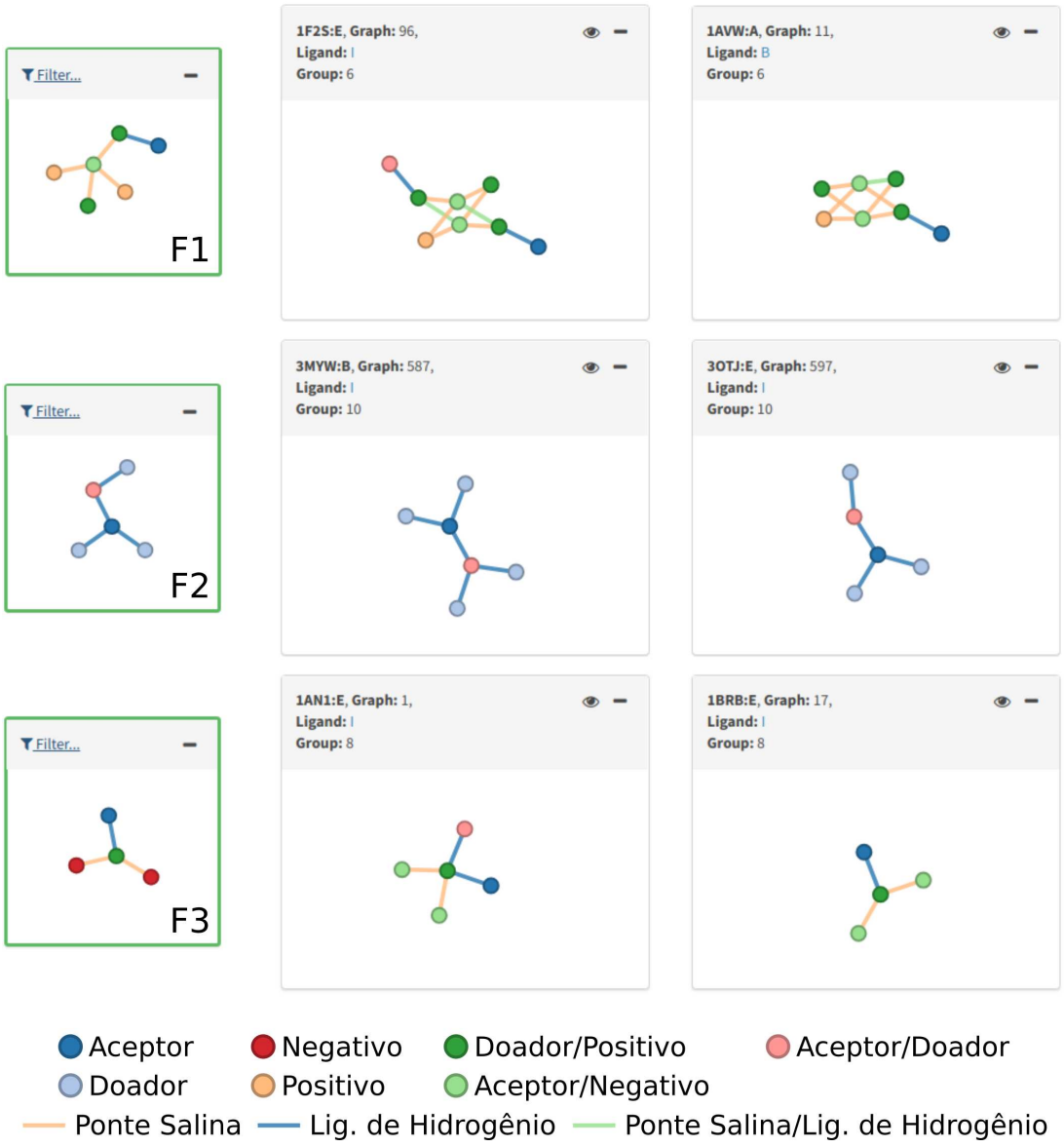


Figura 3.3: Padrões e grafos para conjunto de dados Serino Protease.

O padrão representado por F1 pertence ao grupo 6, e está entre as maiores estruturas conservadas encontradas na base de dados. O vértice central nesta estrutura (vértice com grau 4), rotulado aceptor/negativo, representa nos grafos de origem, um ou ambos os átomos de oxigênio do grupo carboxila presente em resíduos de aspar-

tato, pertencentes a cadeia da protease nos complexos protease/inibidor. Este vértice é encontrado compondo interações do tipo ponte salina com outros quatro vértices da cadeia do inibidor, dos quais dois tem rótulo doador/positivo, e os dois restantes tem rótulo positivo. Nas estruturas onde estes vértices foram encontrados, eles representam os átomos de carbono e nitrogênio do grupo guanidina presente em resíduos de arginina. O último vértice no padrão, rotulado aceptor, encontra-se nos grafos de origem como átomos de oxigênio pertencentes a resíduos de glicina ou serina na cadeia da protease. No padrão, este encontra-se associado a um dos dois vértices rotulados doador/positivo por meio de uma ligação de hidrogênio.

Sob a perspectiva do posicionamento dos resíduos de F1 na estrutura dos complexos protease/inibidor, a maioria dos resíduos de aspartato (presente em 32 grafos) corresponde ao resíduo de aspartato presente no sítio S1 de especificidade em enzimas do tipo tripsina ou tripsina-like. A figura 3.4 mostra o sítio S1 em azul e o resíduo de aspartato correspondente localizado na sua base como Asp189. A única exceção capturada pelos padrões refere-se ao grafo 14, onde o resíduo de aspartato no sítio S1 é deslocado para a posição 226, na estrutura com PDBid 1BRC.

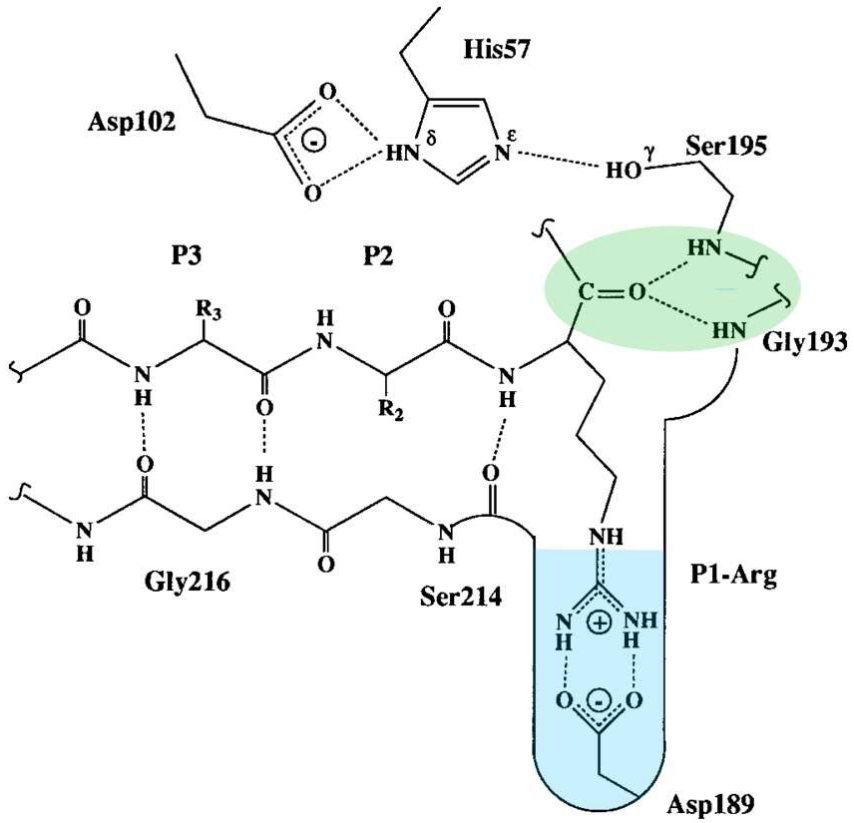


Figura 3.4: Sítio de especificidade S1 em proteínas do tipo tripsina ou tripsina-like. (Fonte: Adaptado de (Perona and Craik, 1995))

Com relação aos resíduos de serina descritos anteriormente, estes estão presentes em trinta dos trinta e três grafos em que F1 foi encontrado, e estão posicionados

imediatamente após o resíduo de aspartato S1, na sequência de resíduos da protease. Já os resíduos de glicina, se localizam mais distantes em relação à sequência, mas próximos espacialmente, devido à conformação da cadeia.

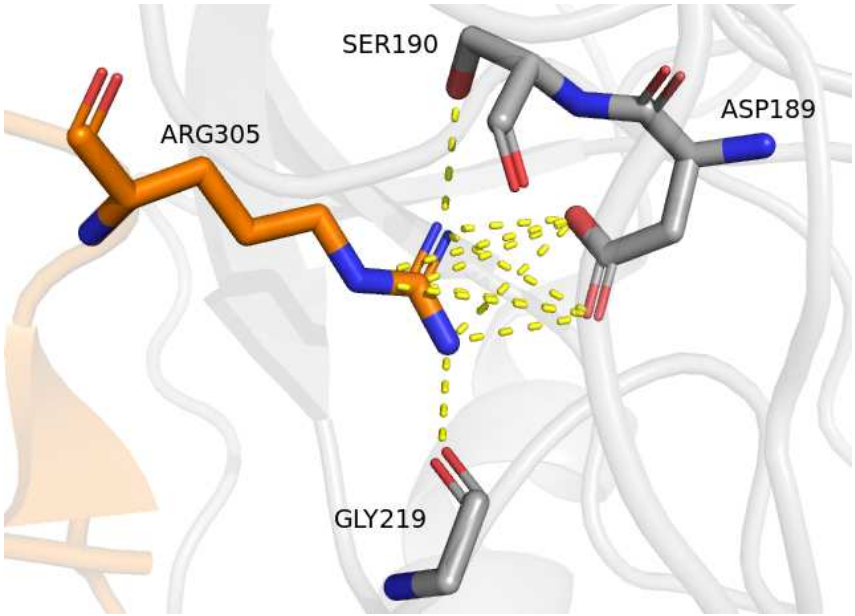


Figura 3.5: Padrão F1 na estrutura com PDBid 1F2S.

A figura 3.5 mostra o padrão F1 no contexto da estrutura cristalina do complexo formado entre uma beta-tripsina bovina e o inibidor de proteínas MCTI-A, PDBid 1F2S.

O padrão F2, representado na figura 3.3, foi extraído do grupo 10 e consiste em cinco vértices, conectados através interações do tipo ligação de hidrogênio. O átomo central nesta estrutura (vértice com grau 3) representa um átomo de oxigênio (aceptor) pertencente ao grupo carboxila de um resíduo da cadeia do inibidor, interagindo com dois átomos de nitrogênio (doadores) e um átomo de oxigênio (aceptor/doador), pertencentes à cadeia da protease. Do lado da protease, um dos átomos de nitrogênio e o átomo de oxigênio correspondem aos átomos de nitrogênio e oxigênio presentes respectivamente no grupo amino e na cadeia lateral de um resíduo de serina, enquanto o outro átomo de nitrogênio no padrão corresponde a um átomo do grupo amino de um resíduo de glicina.

Considerando a posição dos resíduos representados por F2 nos complexos protease/inibidor, os resíduos de serina representam resíduos Ser195 que fazem parte da tríade catalítica das serino-proteases, formada pelos resíduos Ser195, His57 e Asp102. A figura 3.4 ilustra os resíduos da tríade catalítica, e o padrão identificado por F2. A interação (região em verde) nos resíduos de serina e glicina (geralmente Gly193) com o grupo carboxila de outro resíduo, como descrito anteriormente, representa um estado intermediário acil-enzima no processo de catálise dessa classe de enzimas. A figura 3.6 mostra F2 na estrutura do complexo de um inibidor do tipo KTI com uma

tripsina suína, PDBid 4AN7.

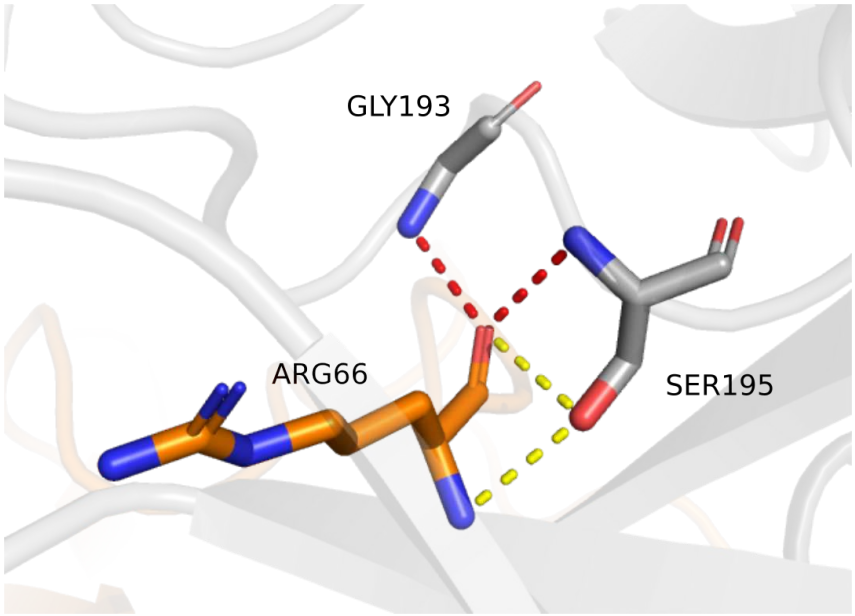


Figura 3.6: Padrão F2 na estrutura PDBid 4AN7.

O último padrão da figura 3.3, identificado como F3, apresenta ambos os átomos de oxigênio (aceptor/negativo) do grupo carboxila presentes em resíduos de aspartato, e um átomo de oxigênio (aceptor ou aceptor/doador) de um resíduo de serina, todos em interação com átomos de nitrogênio de resíduos de lisina pertencentes à cadeia do inibidor. As interações neste padrão representam ligações de hidrogênio e pontes salinas que ocorrem entre o nitrogênio da lisina, respectivamente com os oxigênios das serinas e os oxigênios dos aspartatos descritos anteriormente. O padrão F3 foi obtido a partir do grupo 8, ocorrendo em todos os grafos do grupo.

3.5.2 BCL2

A figura 3.7 mostra exemplos de padrões e alguns grafos onde estes são encontrados.

O grupo 2 é composto de atomos hidrofóbicos e aromaticos conectados por meio de interações do tipo hidrofóbica e ou aromáticas. Um dos padrões encontrados neste grupo, identificado como *F*4, consiste em quatro vértices rotulados hidrofóbicos conectados por interações do hidrofóbicas. O grafo 0 na figura 3.7-A, mostra um grafo onde esse padrão foi encontrado, na estrutura do complexo proteína-peptídio Bcl-xL-Bak, PDBid 1BXL, cadeia A.

Nos grupos 6 e 7, grafos são compostos de vértices rotulados aceptor, aceptor/negativo, doador/positivo, positivo e negativo, conectados por interações do tipo ligação de hidrogênio, ponte salina e repulsiva. Os padrões F5 e F6 os maiores padrões em cada grupo. Uma característica comum nos grafos onde os padrões foram encontrado refere-se à interação de átomos do grupo carboxila em resíduos de aspartato da cadeia

de proteínas anti-apoptóticas, com átomos do grupo guanidina em resíduos de arginina localizados em proteínas pro-apoptóticas. Essas interações foram encontradas predominantemente na forma de pontes salinas. Na figura 3.7-B, vértices representado átomos de aspartato são evidenciados no grafos 39 (complexo Bcl-XL:Beclin 1, PDBid 2P1L, cadeia A) e 401 (estrutura cristalina do foldamero alfa-beta 2c em complexo com Bcl-xL, PDBid 4A1U, cadeia A).

Outra característica em comum nessas estruturas, diz respeito a átomos de resíduos de lisina e arginina na cadeia de proteínas anti-apoptóticas, evidenciadas na figura 3.7-B. Em resíduos de arginina, esse átomos pertencem ao grupo guanidina, enquanto nos resíduos de lisina, estes representam o único átomo de nitrogênio da sua cadeia lateral. Esse átomos são encontrados compondo interações do tipo repulsivas com os átomos dos resíduos arginina da proteína anti-apoptótica descritos anteriormente, e pontes salinas e/ou ligações de hidrogênio com outros átomos das mesmas proteínas.

3.6 Comparação dos padrões com resultados experimentais

3.6.1 Serino protease

Segundo Perona and Craik (1995), o aspartato na posição 189, localizada na base do sítio S1, é altamente conservado em enzimas com especificidade semelhante as tripsinas em relação a substratos que contém resíduos de arginina ou lisina na posição P1 (figura 3.4). O papel do resíduo Asp189, negativamente carregado, no processo de ligação e catálise foi abordado anteriormente em diversos estudos (Graf et al., 1987; Gráf et al., 1988; Perona et al., 1993; Evnin et al., 1990). De acordo com Perona and Craik (1995), a interação nesse resíduo ocorre por meio de ligações de hidrogênio mediadas por uma molécula de água, para substratos contendo resíduos de lisina em P1. Por outro lado, para substratos com resíduos de arginina nessa posição, a interação pode ocorrer também na forma de pontes salinas.

Os padrões encontrados nos grupos 6 e 10, descritos na seção anterior, identificam essas interações predominantemente na forma de pontes salinas, mesmo para resíduos de lisina na posição P1. Isso deve-se principalmente ao critério empregado na prospecção de contatos na região das interfaces, que não considera ligações de hidrogênio mediadas por moléculas de água, mas apenas ligações de hidrogênio formadas diretamente entre átomos do tipo aceptor e doador. Logo, não é esperado encontrar esse tipo de interação. No entanto, considerando o intervalo de distancia similar e as propriedades físico-químicas dos átomos, a estratégia ppiGReMLIN foi capaz de

identificar os respectivos resíduos como relevantes na interface de interação.

Outra interação evidenciada pela estratégia e referenciada na literatura (Perona and Craik (1995)), foi encontrada no grupo 10, onde é representada pelo padrão F2, e no grupo 3, em diversos grafos do grupo. As interações encontradas em ambos os grupos contém uma estrutura conhecida como bolsão oxiânion (*oxyanion hole*), composta por dois átomos de nitrogênio de agrupamentos amida, geralmente proveniente dos resíduos Ser195 e Gly193 das cadeias das proteases, em interação com um átomo de oxigênio proveniente de um grupo carbonila na cadeia do substrato. Essas interações são descritas como do tipo ligação de hidrogênio e podem ser visualizadas nas figuras 3.6 e 3.8, destacadas em vermelho. De acordo com Zakharova et al. (2009), o bolsão oxiânion desempenha seu papel no processo de catálise, estabilizando a carga negativa no grupo carbonila quando este assume um estado intermediário tetraédrico causado pela transição da dupla ligação para um ligação simples entre o átomos de carbono e oxigênio.

Ainda, o resíduo Gly193, altamente conservado em serino-proteases, se mostrou frequente nos grafos descritos acima. De acordo com Bobofchak et al. (2005), sua contribuição é fundamental para a ligação com o substrato à enzima durante os estados de transição e estacionário no processo de catalise.

3.6.2 BCL2

O mecanismo de reconhecimento molecular entre proteínas pro-apoptóticas e anti-apoptóticas da família BCL-2 é norteado pelo domínio BH3, uma cadeia anfipática em alfa-hélice, e sua habilidade de acomodar-se na cavidade formada pelos múltiplos domínios da cadeia dos anti-apoptóticos (figura 3.9-a) (Bhat et al., 2013). Esse modelo de interação é comum às proteínas pro-apoptóticas com único domínio BH3 (ativadores) e com múltiplos domínios (efetores), que atuam na regulação da apoptose através de um mecanismo de competição frente às proteínas anti-apoptóticas (supressores). Em células saudáveis, efetores se encontram complexados, e são substituídos por ativadores como resposta a estímulos pro-apoptose, o que resulta na morte da célula (Dutta et al., 2010).

O domínio BH3 de ativadores e efetores normalmente é composto de vinte resíduos, e caracteriza-se pela presença do trecho de sequencia LXXXXD. Embora interações possam ocorrer em toda a extensão do domínio, apenas alguns resíduos são considerados como críticos para a especificidade de ligação para com os supressores. Além do resíduo de leucina (designado L0), flanqueando o trecho LXXXXD, os resíduos hidrofóbicos nas posições -4, $+3$ e $+7$, representam *hostpots* de ligação no reconhecimento molecular (Bhat et al., 2013). As interações hidrofóbicas promovidas por estes resíduos foram encontradas em padrões nos grupos 1, 2, 3, 4, 5 e 9. O

grafo 0, na figura 3.7-B, destaca vértices representando átomos dos resíduos de leucina (LO) e isoleucina (I+3) em interação com átomos da cadeia anti-apoptótica por meio de interações hidrofóbicas. A estrutura também pode ser visualizada na figura 3.10.

O aspartato D+5, no trecho LXXXXD, é o último dos resíduos considerados críticos na literatura, e interage com átomos de resíduos da cadeia de supressores por meio de ligações de hidrogênio ou pontes salinas (Bhat et al., 2013). Esta interação foi encontrada em grafos nos grupos 6, 7, 8, 10, 11 e 12, como parte de padrões minerados. Os padrões descritos por F5 e F6 na figura 3.7-B, destaca vértices representantes de átomos oriundos de resíduos de aspartato $D + 5$ em interação com átomos do grupo guanidina de resíduos de arginina na cadeia de supressores. A figura 3.11 ilustra as interações representadas pelo grafo de contato 401, relativo ao padrão $F5$, destacando o aspartato $D + 5$ (identificado como Asp121).

Ainda nas estruturas descritas acima, átomos provenientes de resíduos de lisina e arginina na posição $+1$ também foram encontrados compondo interações. Os padrões $F5$ e $F6$ na figura 3.7 destaca esses resíduos como Lys117 no grafo 39 e Arg95 no grafo 401. Embora interações nessa posição não sejam consideradas críticas para a ligação dos domínios BH3 a proteínas anti-apoptóticas, sua contribuição para a afinidade de ligação não pode ser ignorada. Como demonstrado em (Boersma et al., 2008), a substituição de resíduos nessa posição por resíduos de glutamato resultaram na inibição da interação entre a proteína pro-apoptótica Bim-BH3 com as anti-apoptóticas Bcl-xL e Mcl-1. Ainda, Dutta et al. (2010) mostrou diferentes paradigmas de interação para o mesmo conjunto de proteínas, onde a substituição por resíduos com carga negativa em Bim-BH3 resultou em baixa afinidade de ligação para com Bcl-xL, sendo mais toleradas no caso de Mcl-1.

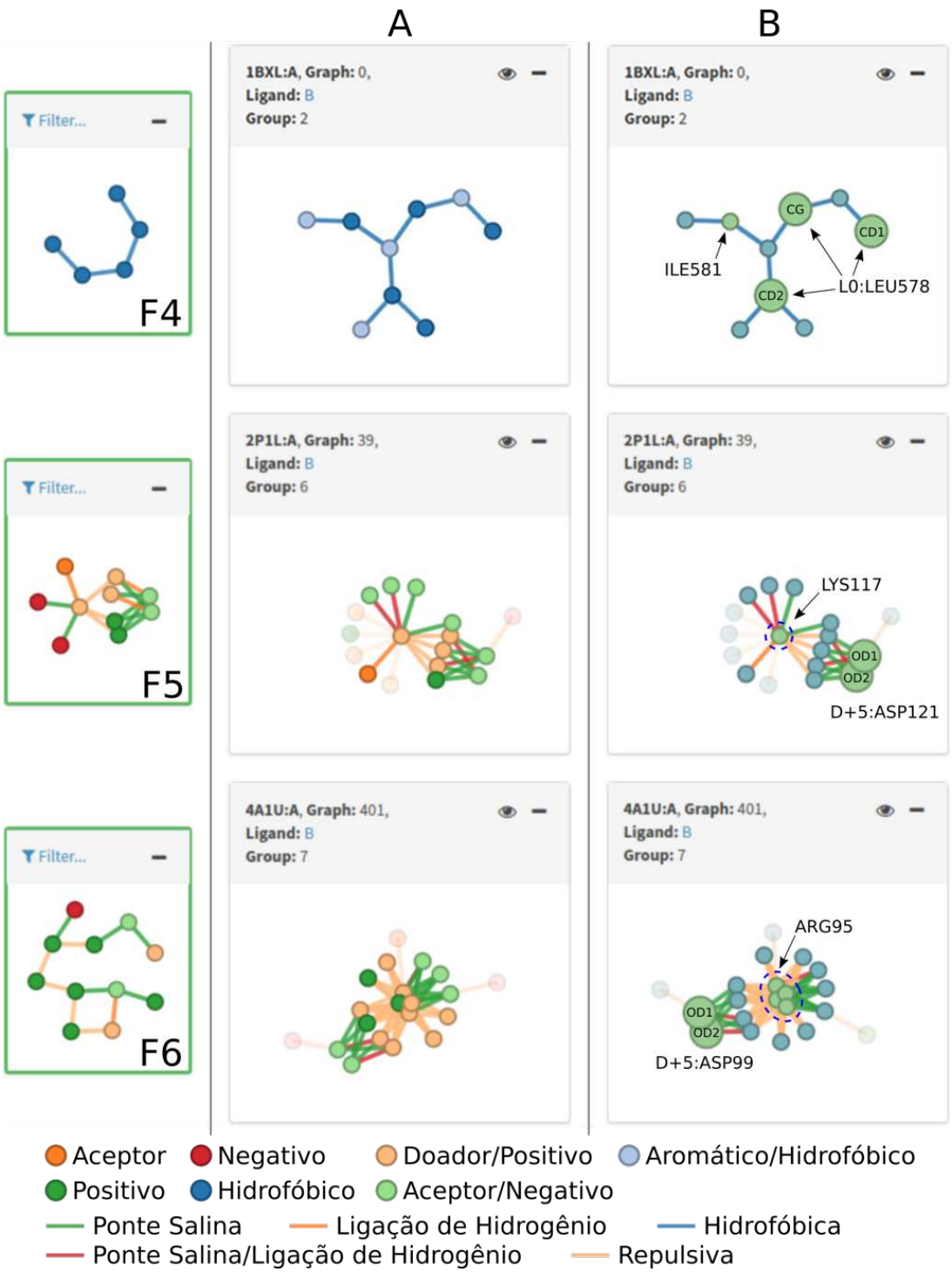


Figura 3.7: Padrões e grafos para conjunto de dados BCL2.

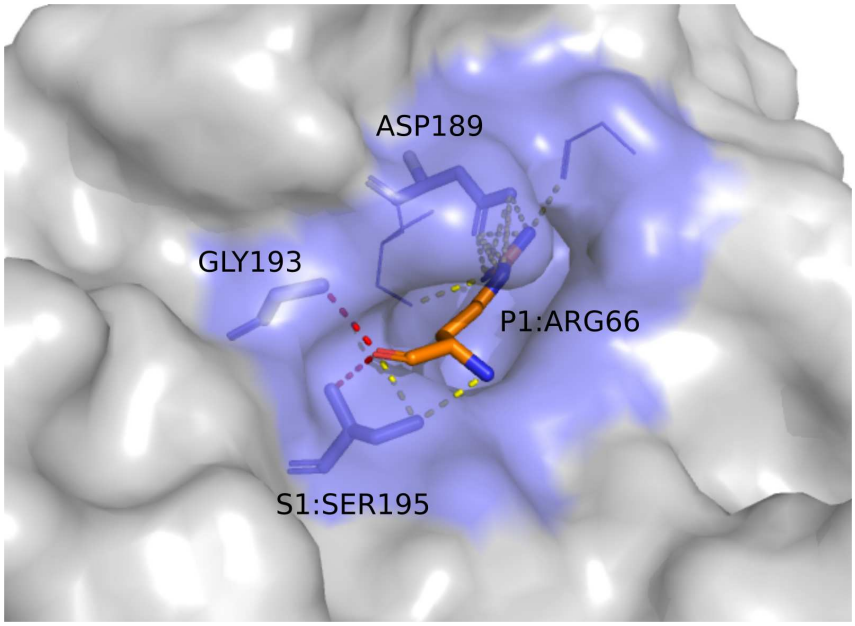
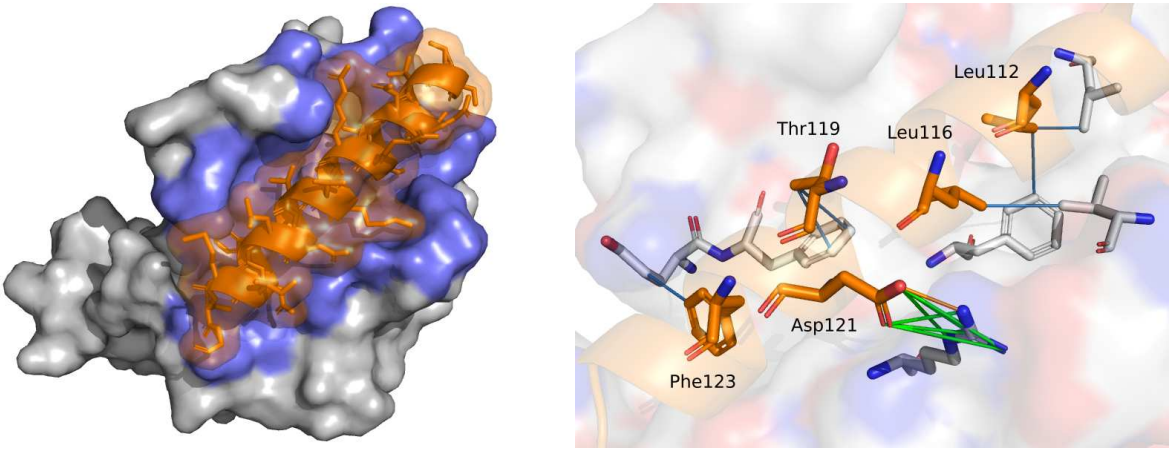


Figura 3.8: Grafos com padrões F1 e F2 mostrados no sítio de ligação S1, e resíduo P1 (ARG66) na estrutura com PDBid 4AN7.



(a) (b)

Figura 3.9: Interação BH3/BCL2 na estrutura com PDBid 2P1L. a) Dominio BH3 na cavidade de interação BCL2 b) Resíduos críticos na interação

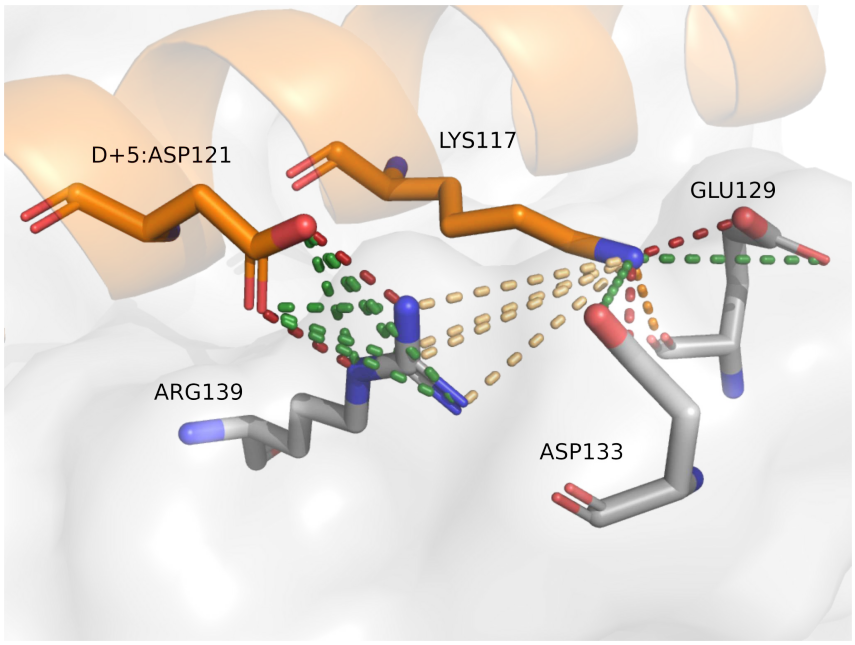


Figura 3.10: Grafo 39, contendo padrão F6, representado na estrutura 4A1U

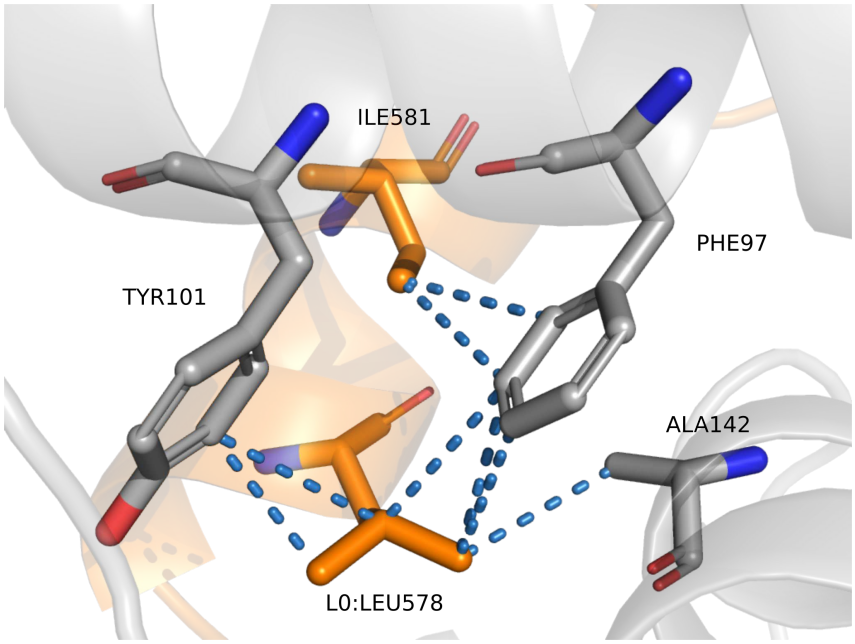


Figura 3.11: Grafo 0, contendo padrão F4, na estrutura com PDBid 1BXL

Capítulo 4

Conclusões e Perspectivas

Esta trabalho propôs ppiGReMLIN, uma estratégia para detecção de estruturas conservadas em interfaces proteína-proteína baseada em grafos. Para isso as interfaces proteína-proteína são representadas como grafos de interações no nível atômico, onde vértices e arestas correspondem respectivamente, aos átomos presentes na interface e as interações não-covalentes entre estes, que são inferidas segundo critérios de distancia entre os respectivos átomos e suas propriedades físico-químicas. O conjunto de grafos é dividido em grupos, de onde estruturas conservadas são obtidas pela aplicação de técnicas de mineração de subgrafos frequentes, revelando estruturas conservadas para todo o conjunto de dados. O método proposto não depende de alinhamento de sequência ou superposição de estruturas, e pode ser aplicado para conjuntos de dados de interações proteína-proteína em larga escala.

Os resultados mostram a efetividade do método em identificar estruturas conservadas descritas na literatura de maneira automática. Para o conjunto de dados de serino proteases, por exemplo, padrões foram encontrados com suporte mínimo superior a 0.7. Estruturas relevantes na interação entre a protease e seus substrato foram encontradas, como o oxiânion envolvendo os resíduos Ser195 e Gly193, importante na formação dos estados intermediários do processo de catálise, e o resíduo Asp189, localizado no sítio de especificidade das tripsinas, que foram detectadas em múltiplos grupos.

Para o conjunto de dados BCL-2/BH3, estruturas relevantes foram encontradas com suporte mínimo superior a 0.6. Nesse caso, resíduos em posições críticas relativas ao trecho de sequencia LXXXXD no domínio BH3 de proteínas pro-apoptóticas foram encontradas em diversos padrões, em diferentes grupos. Além disso, resíduos relevantes também foram encontradas em outras posições, que apesar de serem consideradas como não criticas, influenciam a afinidade de ligação entre reguladores e supressores da família BCL-2

Como trabalhos futuros, inicialmente pretende-se estender os critérios de prospecção de contatos nas interfaces proteína-proteína visando extrair outros tipos de interações, como ligações de hidrogênio mediadas por moléculas de água e pontes dissulfeto. Pretende-se também disponibilizar a estratégia ppiGReMLIN como *web*

server, de modo a tornar o método acessível para usuários, e como *web service*, podendo ser incluído como parte integrante de outros processos automatizados.

Referências Bibliográficas

(2013). Genbank: Accession no. jx898746.1.

(2017). Embrapa. https://www.embrapa.br/soja/cultivos/soja1/dados-economicos. [Online; acessado 1-Novembro-2017].

(2017). OEC. https://atlas.media.mit.edu/pt/profile/country/bra/. [Online; acessado 1-Novembro-2017].

Aggarwal, C. C. (2015). *Data mining: the textbook*. Springer.

Alberts, B. (2017). *Biologia molecular da célula*. Artmed, 6 edition.

Bhat, V., Olenick, M. B., Schuchardt, B. J., Mikles, D. C., McDonald, C. B., and Farooq, A. (2013). Biophysical basis of the promiscuous binding of b-cell lymphoma protein 2 apoptotic repressor to bh3 ligands. *Journal of Molecular Recognition*, 26(10):501–513.

Bobofchak, K. M., Pineda, A. O., Mathews, F. S., and Di Cera, E. (2005). Energetic and structural consequences of perturbing gly-193 in the oxyanion hole of serine proteases. *Journal of Biological Chemistry*, 280(27):25644–25650.

Boersma, M. D., Sadowsky, J. D., Tomita, Y. A., and Gellman, S. H. (2008). Hydrophile scanning as a complement to alanine scanning for exploring and manipulating protein–protein recognition: application to the bim bh3 domain. *Protein Science*, 17(7):1232–1240.

Bondy, A. and Murty, U. (2011). *Graph Theory*. Graduate Texts in Mathematics. Springer London.

Borgelt, C. and Berthold, M. R. (2002). Mining molecular fragments: Finding relevant substructures of molecules. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 51–58. IEEE.

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

da Silveira, C. H., Pires, D. E., Minardi, R. C., Ribeiro, C., Veloso, C. J., Lopes, J. C., Meira, W., Neshich, G., Ramos, C. H., Habesch, R., et al. (2009). Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for

prospecting contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, 74(3):727–743.

Delbridge, A. and Strasser, A. (2015). The bcl-2 protein family, bh3-mimetics and cancer therapy. *Cell death and differentiation*, 22(7):1071.

Delbridge, A. R., Grabow, S., Strasser, A., and Vaux, D. L. (2016). Thirty years of bcl-2: translating cell death discoveries into novel cancer therapies. *Nature reviews Cancer*, 16(2):99.

Delbridge, A. R., Valente, L. J., and Strasser, A. (2012). The role of the apoptotic machinery in tumor suppression. *Cold Spring Harbor perspectives in biology*, 4(11):a008789.

Dutta, S., Gullá, S., Chen, T. S., Fire, E., Grant, R. A., and Keating, A. E. (2010). Determinants of bh3 binding specificity for mcl-1 versus bcl-xl. *Journal of molecular biology*, 398(5):747–762.

Eldén, L. (2006). Numerical linear algebra in data mining. *Acta Numerica*, 15:327–384.

Elmore, S. (2007). Apoptosis: a review of programmed cell death. *Toxicologic pathology*, 35(4):495–516.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

Evnin, L. B., Vásquez, J. R., and Craik, C. S. (1990). Substrate specificity of trypsin investigated by using a genetic selection. *Proceedings of the National Academy of Sciences*, 87(17):6659–6663.

Fassio, A. V., Martins, P. M., Guimarães, S. d. S., Junior, S. S., Ribeiro, V. S., de Melo-Minardi, R. C., and Silveira, S. d. A. (2017). Vermont: a multi-perspective visual interactive platform for mutational analysis. *BMC bioinformatics*, 18(10):403.

Fassio, A. V., Santana, C. A., Cerqueira, F. R., da Silveira, C. H., Romanelli, J. P., de Melo-Minardi, R. C., and Silveira, S. d. A. (2018). An interactive strategy to visualize common subgraphs in protein-ligand interaction. In *International Conference on Bioinformatics and Biomedical Engineering*, pages 383–394. Springer.

Fromm, H. and Hargrove, M. (2011). *Essentials of Biochemistry*. SpringerLink : Bücher. Springer Berlin Heidelberg.

Gonçalves-Almeida, V. M., Pires, D. E., de Melo-Minardi, R. C., da Silveira, C. H., Meira, W., and Santoro, M. M. (2011). Hydropace: understanding and predicting

cross-inhibition in serine proteases through hydrophobic patch centroids. *Bioinformatics*, 28(3):342–349.

Graf, L., Craik, C. S., Patthy, A., Roczniak, S., Fletterick, R. J., and Rutter, W. J. (1987). Selective alteration of substrate specificity by replacement of aspartic acid-189 with lysine in the binding pocket of trypsin. *Biochemistry*, 26(9):2616–2623.

Gráf, L., Jancso, A., Szilágyi, L., Hegyi, G., Pintér, K., Náray-Szabó, G., Hepp, J., Medzihradszky, K., and Rutter, W. J. (1988). Electrostatic complementarity within the substrate-binding pocket of trypsin. *Proceedings of the National Academy of Sciences*, 85(14):4961–4965.

Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using networkx. In Varoquaux, G., Vaught, T., and Millman, J., editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA.

Han, J., Kamber, M., and Pei, J. (2006). Data mining: Concepts and techniques. pages 585–631.

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

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *cell*, 144(5):646–674.

Harary, F. (1969). *Graph theory*. Addison-Wesley series in mathematics. Addison-Wesley Pub. Co., Reading.

Hinneburg, A., Keim, D. A., et al. (1998). An efficient approach to clustering in large multimedia databases with noise. In *KDD*, volume 98, pages 58–65.

Hsieh, S.-M., Hsu, C.-C., and Hsu, L.-F. (2006). Efficient method to perform isomorphism testing of labeled graphs. In *International Conference on Computational Science and Its Applications*, pages 422–431. Springer.

Huan, J., Wang, W., Prins, J., and 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*, pages 581–586. ACM.

Jiang, C., Coenen, F., and Zito, M. (2013). A survey of frequent subgraph mining algorithms. *The Knowledge Engineering Review*, 28(1):75–105.

Khashan, R., Zheng, W., and Tropsha, A. (2012). Scoring protein interaction decoys using exposed residues (spider): a novel multibody interaction scoring function

based on frequent geometric patterns of interfacial residues. *Proteins: Structure, Function, and Bioinformatics*, 80(9):2207–2217.

Koyutürk, M., Grama, A., and Szpankowski, W. (2004). An efficient algorithm for detecting frequent subgraphs in biological networks. *Bioinformatics*, 20(suppl_1):i200–i207.

Kuramochi, M. and Karypis, G. (2004). An efficient algorithm for discovering frequent subgraphs. *IEEE Transactions on Knowledge and Data Engineering*, 16(9):1038–1051.

Leskovec, J., Rajaraman, A., and Ullman, J. (2014). Mining of massive datasets, chapter 11: Dimensionality reduction.

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

Moreira, I. S., Fernandes, P. A., and Ramos, M. J. (2007). Hot spots—a review of the protein–protein interface determinant amino-acid residues. *Proteins: Structure, Function, and Bioinformatics*, 68(4):803–812.

Morozova, N., Allers, J., Myers, J., and Shamoo, Y. (2006). Protein–rna interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution structures. *Bioinformatics*, 22(22):2746–2752.

Mrzic, A., Meysman, P., Bittremieux, W., Moris, P., Cule, B., Goethals, B., and Laukens, K. (2018). Grasping frequent subgraph mining for bioinformatics applications. *BioData mining*, 11(1):20.

Nelson, D. and Cox, M. (2018). *Princípios de Bioquímica de Lehninger - 7.ed.* Artmed Editora.

Nelson, D. L., Lehninger, A. L., and Cox, M. M. (2008). *Lehninger principles of biochemistry*. Macmillan.

Nevola, L. and Giralt, E. (2015). Modulating protein–protein interactions: the potential of peptides. *Chemical Communications*, 51(16):3302–3315.

Ofran, Y. and Rost, B. (2003). Analysing six types of protein–protein interfaces. *Journal of molecular biology*, 325(2):377–387.

Patarroyo-Vargas, A. M., Merino-Cabrera, Y. B., Zanuncio, J. C., Rocha, F., Campos, W. G., de Almeida, O., and Maria, G. (2017). Kinetic characterization of anticarsia gemmatalis digestive serine-proteases and the inhibitory effect of synthetic peptides. *Protein and peptide letters*, 24(11):1040–1047.

Perona, J. J. and Craik, C. S. (1995). Structural basis of substrate specificity in the serine proteases. *Protein Science*, 4(3):337–360.

Perona, J. J., Tsu, C. A., McGrath, M. E., Craik, C. S., and Fletterick, R. J. (1993). Relocating a negative charge in the binding pocket of trypsin. *Journal of molecular biology*, 230(3):934–949.

Petros, A. M., Olejniczak, E. T., and Fesik, S. W. (2004). Structural biology of the bcl-2 family of proteins. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1644(2-3):83–94.

Pilon, F., Visôtto, L., Guedes, R., and Oliveira, M. (2013). Proteolytic activity of gut bacteria isolated from the velvet bean caterpillar anticarsia gemmatalis. *Journal of Comparative Physiology B*, 183(6):735–747.

Pilon, F. M., Silva, C. d. R., Visôtto, L. E., Barros, R. d. A., da Silva Júnior, N. R., Campos, W. G., and de Almeida Oliveira, M. G. (2017). Purification and characterization of trypsin produced by gut bacteria from anticarsia gemmatalis. *Archives of insect biochemistry and physiology*, 96(2):e21407.

Saidi, R., Maddouri, M., and 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*, pages 35–38. ACM.

Santana, C. A., Cerqueira, F. R., da Silveira, C. H., Fassio, A. V., de Melo-Minardi, R. C., and Silveira, S. d. A. (2016). Gremlin: A graph mining strategy to infer protein-ligand interaction patterns. In *Bioinformatics and Bioengineering (BIBE), 2016 IEEE 16th International Conference on*, pages 28–35. IEEE.

Scott, D. E., Bayly, A. R., Abell, C., and Skidmore, J. (2016). Small molecules, big targets: drug discovery faces the protein–protein interaction challenge. *Nature Reviews Drug Discovery*, 15(8):533.

Scott, I. M., Thaler, J. S., and Scott, J. G. (2010). Response of a generalist herbivore trichoplusia ni to jasmonate-mediated induced defense in tomato. *Journal of chemical ecology*, 36(5):490–499.

Shirinivas, S., Vetrivel, S., and Elango, N. (2010). Applications of graph theory in computer science an overview. *International Journal of Engineering Science and Technology*, 2(9):4610–4621.

Silva, F., Oliveira, M. d. A., Batista, R., Pires, C., Xavier, L., Piovesan, N., Oliveira, J., José, I., and Moreira, M. (2002). Função fisiológica de lipoxigenases de folhas de

soja submetidas ao ataque de lagarta (anticarsia gemmatalis hübner). *Arquivo do Instituto Biológico*, 69(1):67–74.

Silveira, S. A., Fassio, A. V., Gonçalves-Almeida, V. M., de Lima, E. B., Barcelos, Y. T., Aburjaile, F. F., Rodrigues, L. M., Meira Jr, W., and de Melo-Minardi, R. C. (2014). Vermont: Visualizing mutations and their effects on protein physicochemical and topological property conservation. In *BMC proceedings*, volume 8, page S4. BioMed Central.

Sorenson, C. M. (2004). Bcl-2 family members and disease. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1644(2-3):169–177.

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

Terra, W. R. and Ferreira, C. (1994). Insect digestive enzymes: properties, compartmentalization and function. *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry*, 109(1):1–62.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

Wielkopolan, B., Walczak, F., Podleśny, A., Nawrot, R., and Obrępalska-Stęplowska, A. (2015). Identification and partial characterization of proteases in larval preparations of the cereal leaf beetle (oulema melanopus, chrysomelidae, coleoptera). *Archives of insect biochemistry and physiology*, 88(3):192–202.

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

Zakharova, E., Horvath, M. P., and Goldenberg, D. P. (2009). Structure of a serine protease poised to resynthesize a peptide bond. *Proceedings of the National Academy of Sciences*, 106(27):11034–11039.

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